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

Verso un supporto (AI)ntelligente per gli utenti del Tier-1

M. Barbetti*, D. Cesini, C. Giugliano, D. Lattanzio, L. Morganti, A. Pascolini
A. Rendina, E. Ronchieri, A. Shtimmerman, A. Trashaj, C. Pellegrino

Workshop sul Calcolo nell'INFN | 20-24 maggio 2024

WLCG Tier-1 made in Italy

Credits: Pier Paolo Ricci (INFN CNAF)

Il Tier-1 italiano della WLCG si trova a **Bologna**

- trasferimento ancora in corso verso il **Tecnopolo** (ex Manifattura Tabacchi)
- gestito dal **CNAF** fin dal 2003 [1]

RISORSE

~2k nodi di calcolo (fisici e virtuali)
- **~55k core** (20k + 35k @ CNAF/CINECA)
- **~670 kHS06** (265 kHS06 + 405 kHS06 @ CNAF/CINECA)
- risorse gestite da un ***batch system*** **centralizzato** (HTCondor)

~70 PB di disco (per conservare i "dati caldi")

~130 PB di nastro (per conservare i "dati freddi")

Il nuovo Tier-1 al Tecnopolo subirà un **importante aggiornamento** [2] in vista di HL-LHC

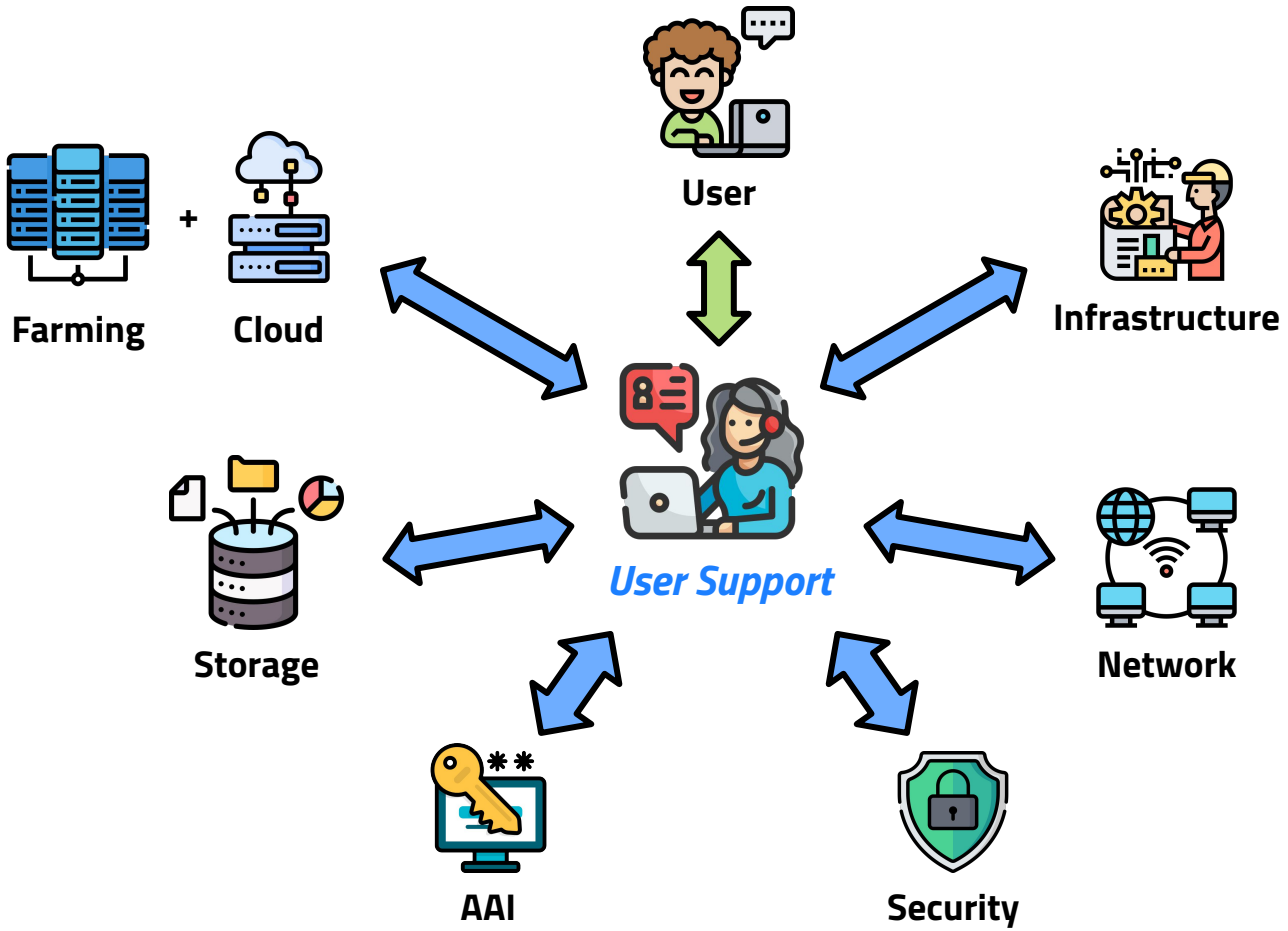
- **CPU:** ~670 kHS06 (2023) → **~1140 kHS06** (2026)
- **Disco:** ~70 PB (2023) → **~120 PB** (2026)
- **Nastro:** ~130 PB (2023) → **~280 PB** (2026)

Il calcolo scientifico al Tier-1

Il Tier-1 fornisce risorse per il calcolo scientifico a un'**ampia comunità scientifica** [1]

- collaborazione con **60+ esperimenti** (non solo i quattro esperimenti principali di LHC)
- supporto ai **1500+ utenti attivi** aventi accesso alle risorse (non tutti gli utenti operano nel campo della fisica)

Come supportare gli utenti/esperimenti

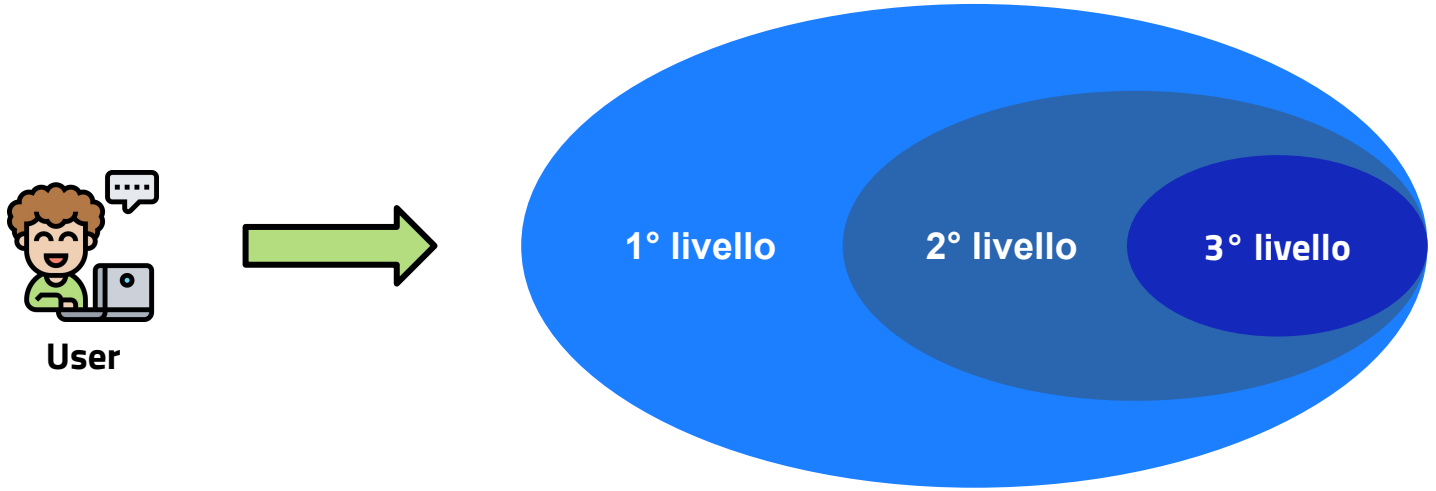


Il Tier-1 garantisce supporto agli esperimenti/utenti tramite il reparto dedicato ***User Support*** (US) [3] che

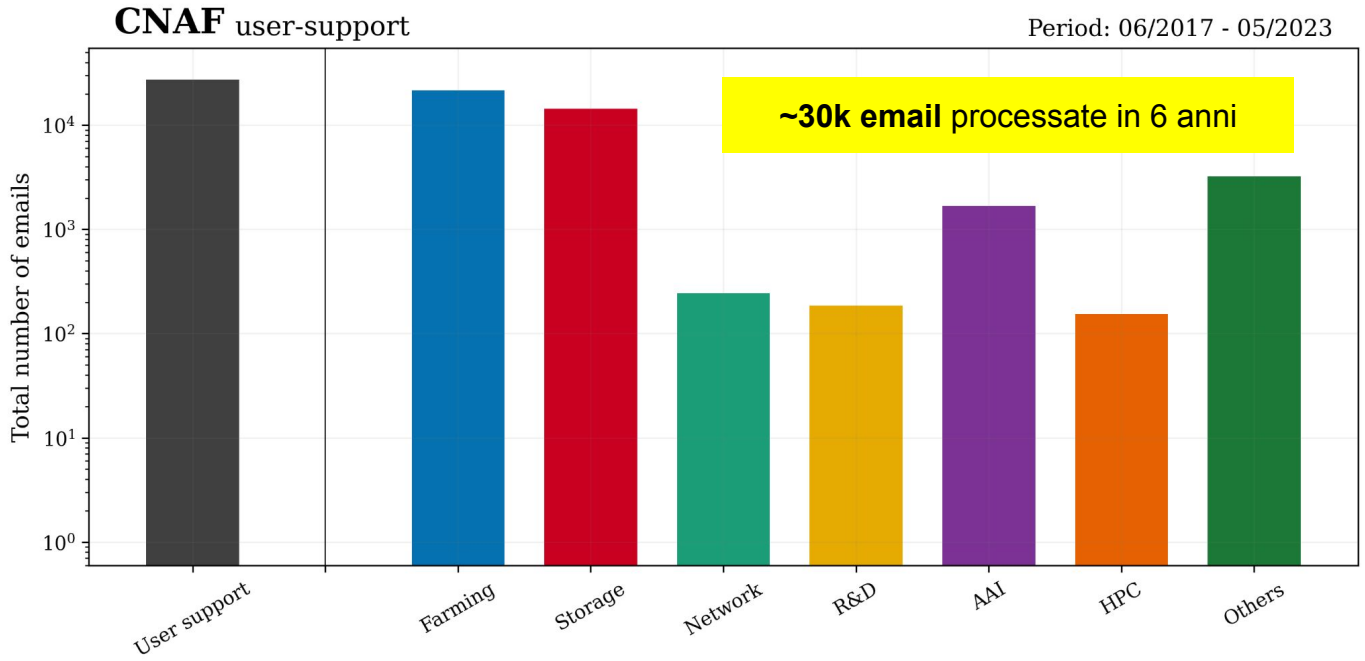
- agisce come **principale punto di contatto** tra gli utenti e i reparti specializzati del Tier-1
- sviluppa strumenti e procedure per **semplificare** l'accesso alle risorse da parte degli utenti
- assiste gli utenti per un **utilizzo efficace ed efficiente** delle risorse di calcolo, disco e nastro
- collabora con i diversi esperimenti per definire un ***computing model*** conforme agli standard del Tier-1
- prepara e tiene aggiornata la **documentazione ufficiale** per gli utenti del Tier-1 (https://l.infn.it/t1guide)

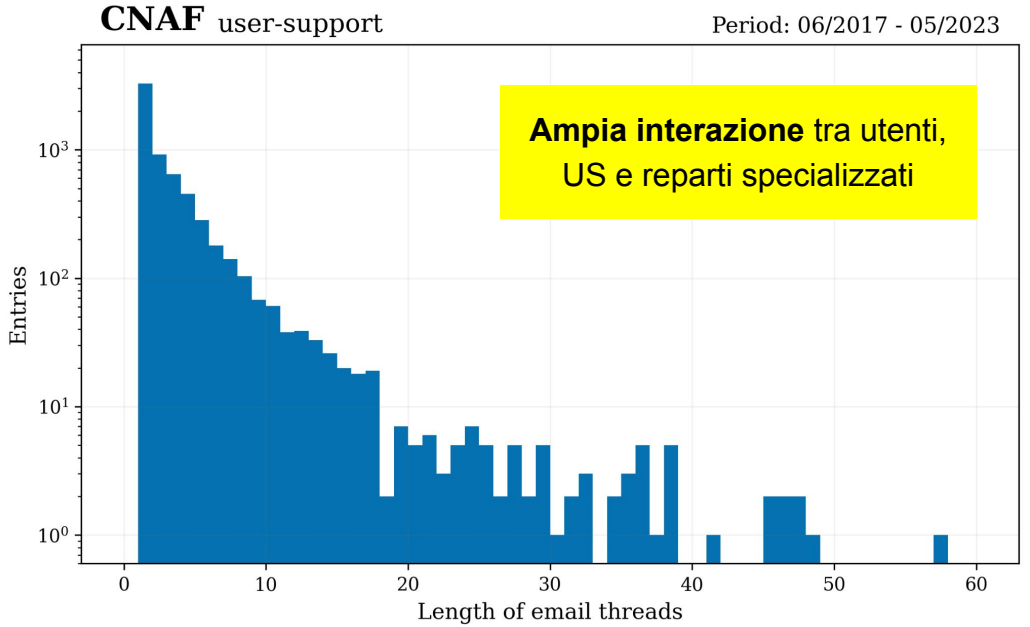
La missione del reparto User Support

- Alto numero di utenti + sviluppo/adozione di nuove tecnologie → **il ruolo del reparto US è cruciale**
- Utenti proveniente da comunità scientifiche diverse → **esigenze diverse** per il calcolo negli esperimenti
- Supporto a livelli → **1° livello (*User Support*)**, 2° livello (reparto specializzato), 3° livello (sviluppatore software)



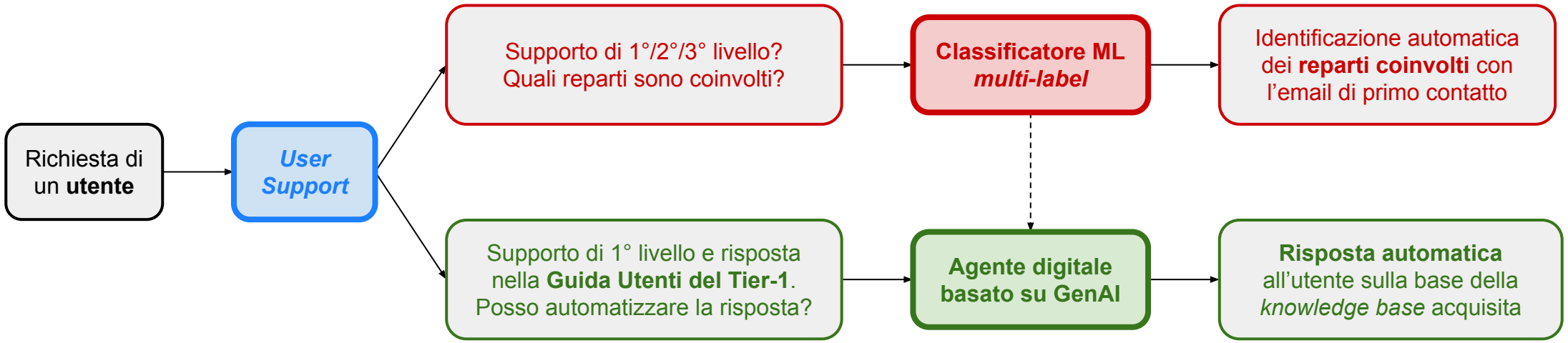
Un po' di metriche





Evolvere lo User Support con l'AI

- Aggiornamento del Tier-1 → **nuovi utenti**, **nuovi esperimenti** (potenzialmente) da comunità scientifiche diverse e differenti
- Supportare il nuovo bacino di utenti richiede un'**evoluzione** delle procedure attualmente in uso, incluso il reparto US
- Tecniche di ***Machine Learning*** (ML) all'avanguardia e moderni **Modelli di Linguaggio** (basati su **GenAI**) possono aiutare nel supporto di un crescente numero di utenti e nell'adozione delle ultime tecnologie software → come? [4]



CLASSIFICATORE MULTI-LABEL

Ricetta per la costruzione del classificatore

1. Preparazione del **campione di email** per l'addestramento
2. Trasformazione delle **feature del testo** in **rappresentazione numerica**
3. Addestramento di vari **modelli di classificazione**
4. Misura delle **performance** di classificazione e **combinazione** dei risultati

Preparazione del dataset di addestramento

Raccolta dei dati

Campione di **~30k email** ricevute/inviate nel periodo 06/2017 - 05/2023

Email salvate come singoli file JSON (**~260 MB**)

No allegati, (quasi) no HTML

Ciascun file include:

- `from` – indirizzo email
- `to` – indirizzo email
- `date` – data e ora
- `subject` – testo
- `content` – testo
- `parent` – indirizzo email

Anonimizzazione

Tutti gli **indirizzi email** sono stati **anonimizzati** sostituendoli con degli UUID

Tutti i **riferimenti a persone** (es.: nomi propri, *username*) contenuti nell'oggetto/corpo delle email sono stati **anonimizzati** e sostituiti con il *placeholder* `[NAME]`

L'anonimizzazione è eseguita mediante **script automatici** e **modifiche manuali** ai *metadata* delle email

Labelling

A ciascuna email viene associata un'**etichetta** (*label*) in base al testo contenuto nell'oggetto/corpo

Il ***thread* di email** tra utente e i vari reparti del Tier-1 viene ricostruito sfruttando le info contenute in `parent`

Tutte le etichette delle email di uno stesso *thread* vengono propagate fino all'**email di primo contatto** (***labelling***)

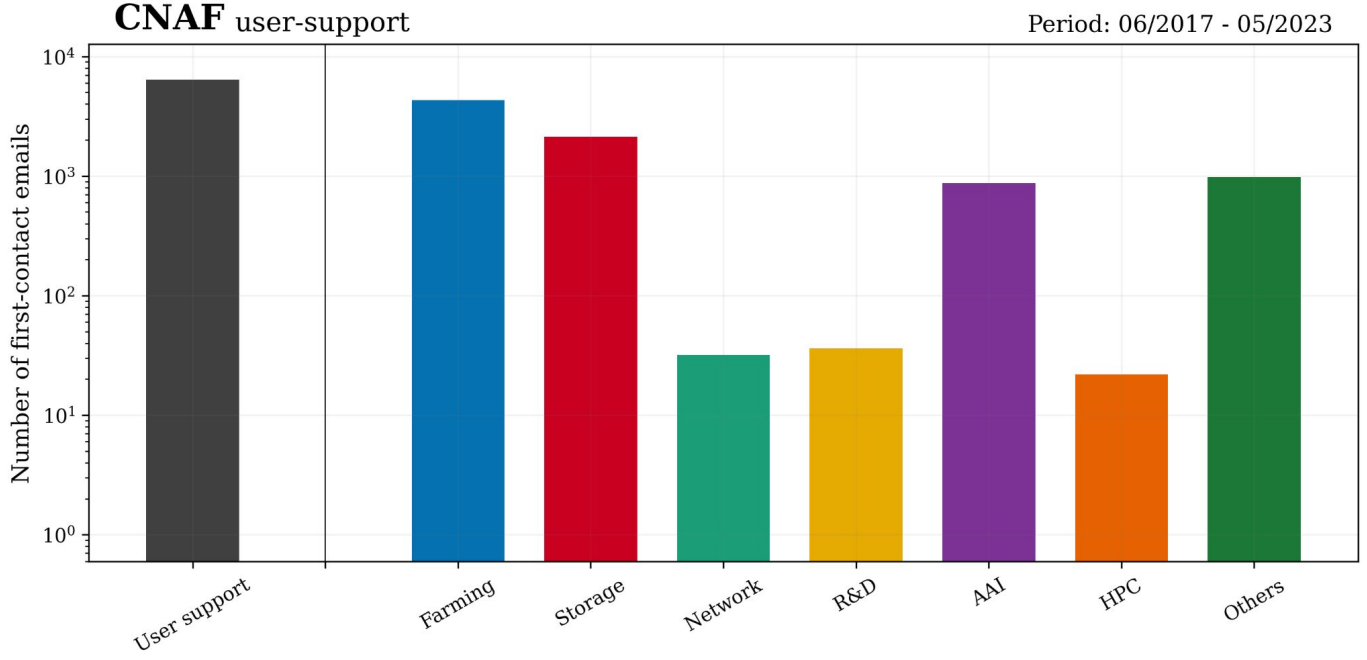
Data cleaning

Il corpo delle email può contenere ***escape sequence*** (es.: `\n`, `\t`) o **tag HTML**

Le email sono **multilingua** (inglese/italiano) e possono contenere **caratteri speciali** (es.: lettere accentate, emoji)

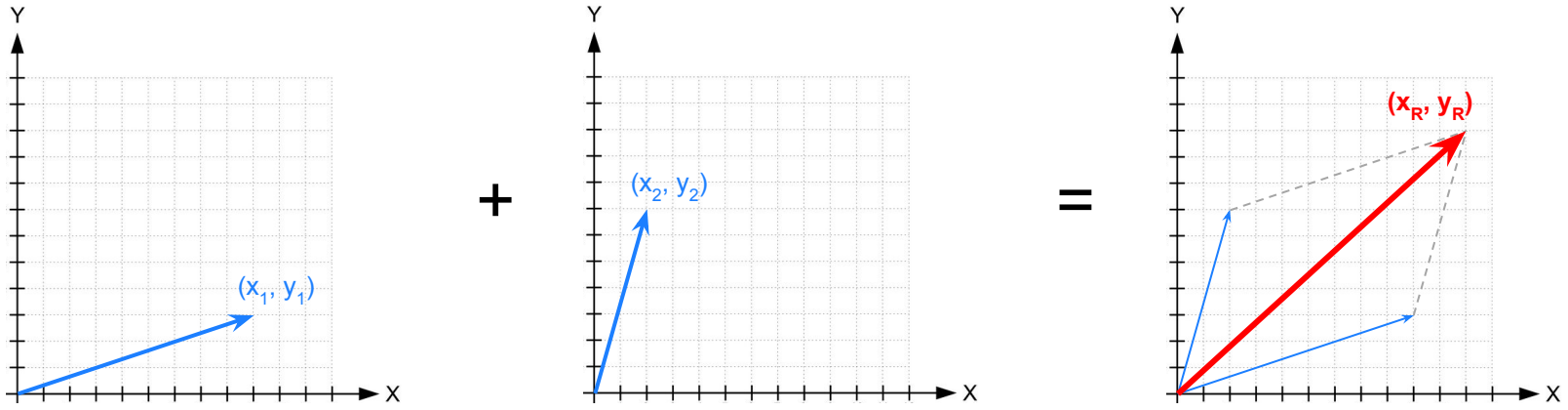
La procedura di ***data cleaning*** trasforma il testo delle email in modo che sia composto solo da parole, *placeholder* e punteggiatura

Il dataset di addestramento



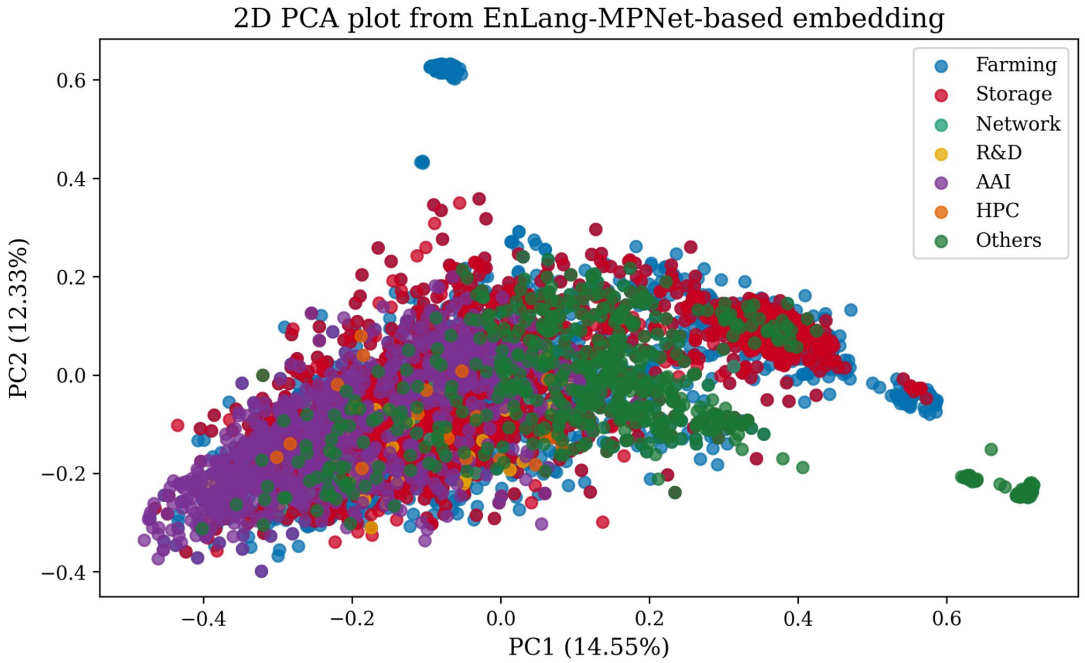
Rappresentazione delle feature testuali

- Il significato di un testo viene codificato in combinazioni di parole → i modelli di ML lavorano con i vettori (***feature***)
  - Combinazioni di parole (o *token*) possono essere mappate in uno **"spazio di rappresentazione"** (spazio vettoriale ad alta dimensionalità) tramite un **modello di *embedding***
- Il modello di embedding può essere costruito in modo che rispetti la **semantica**, anche in molteplici lingue
  - Sentence-Transformers, framework Python che fornisce modelli aggiornati per l'*embedding* del testo



Rappresentazione di "👩" + Rappresentazione di "👑" = Rappresentazione di "👸"

Esempio di email nello spazio di rappresentazione



`all-mpnet-base-v2`

- **modello base:** MPNet
- **dimensione *embedding*:** 768
- **max num di token:** 384
- **velocità inferenza:** 1
- **lingua:** solo inglese

`paraphrase-multilingual-mpnet-base-v2`

- **modello base:** XLM-RoBERTa
- **dimensione *embedding*:** 768
- **max num di token:** 128
- **velocità inferenza:** ~1
- **lingua:** multipla

`all-MiniLM-L6-v2`

- **modello base:** MiniLM
- **dimensione *embedding*:** 384
- **max num di token:** 256
- **velocità inferenza:** x5
- **lingua:** solo inglese

`paraphrase-multilingual-MiniLM-L12-v2`

- **modello base:** MiniLM
- **dimensione *embedding*:** 384
- **max num di token:** 126
- **velocità inferenza:** x2.5
- **lingua:** multipla

Performance dei modelli addestrati

Diversi modelli ML sono stati addestrati per eseguire una **classificazione *multi-label***

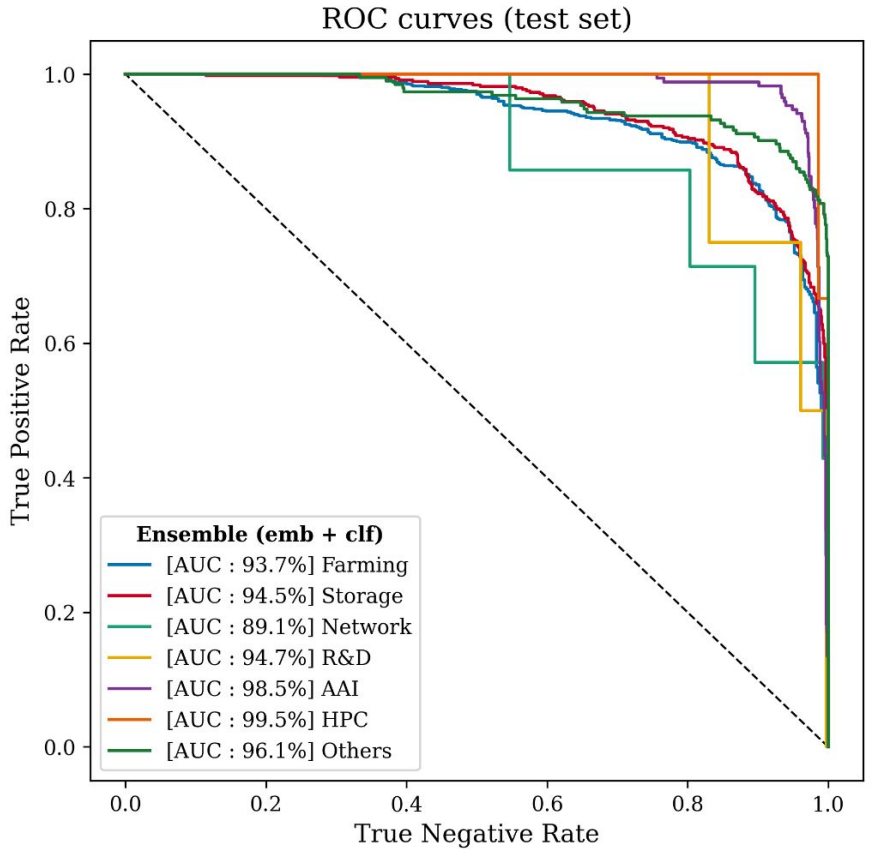
- ***k-Nearest Neighbors*** (kNN)
- ***Random Forest*** (RF)
- ***Extreme Gradient Boosting*** (XGBoost)
- ***Feed-forward Neural Network*** (FNN)

Diversi *embedding* sono stati usati in combinazione con i vari modelli di classificazione

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

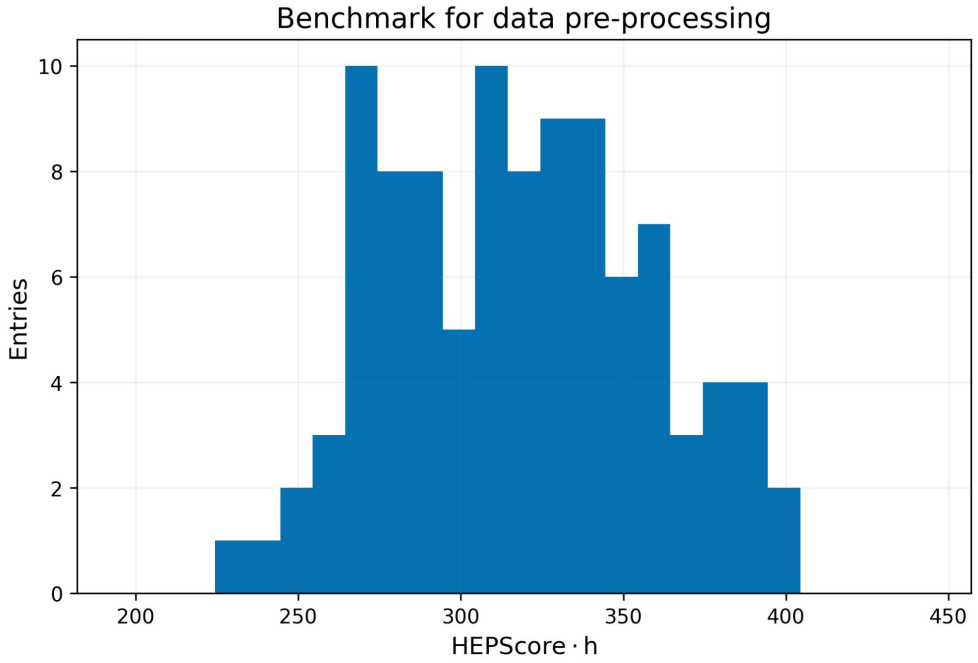
Diverse strategie di *preprocessing* → **4 (emb) x 4 (prep) x 4 (clf) = 64 modelli diversi**

Combinare "opportunamente" l'output dei modelli più promettenti (***ensemble***) permette di ottenere le *performance* migliori: **~95% di ROC AUC** nel campione di test [4]

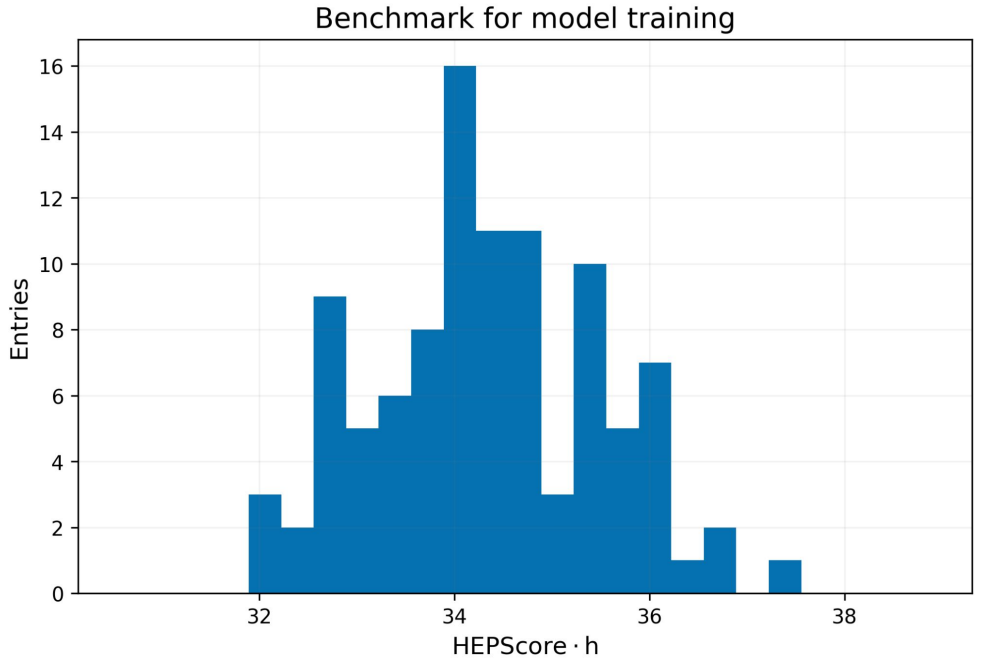


Costo computazionale per l'addestramento del classificatore

Studi di *performance* temporali eseguiti sui **nodi di calcolo** della *farm* del Tier-1 (no GPU)



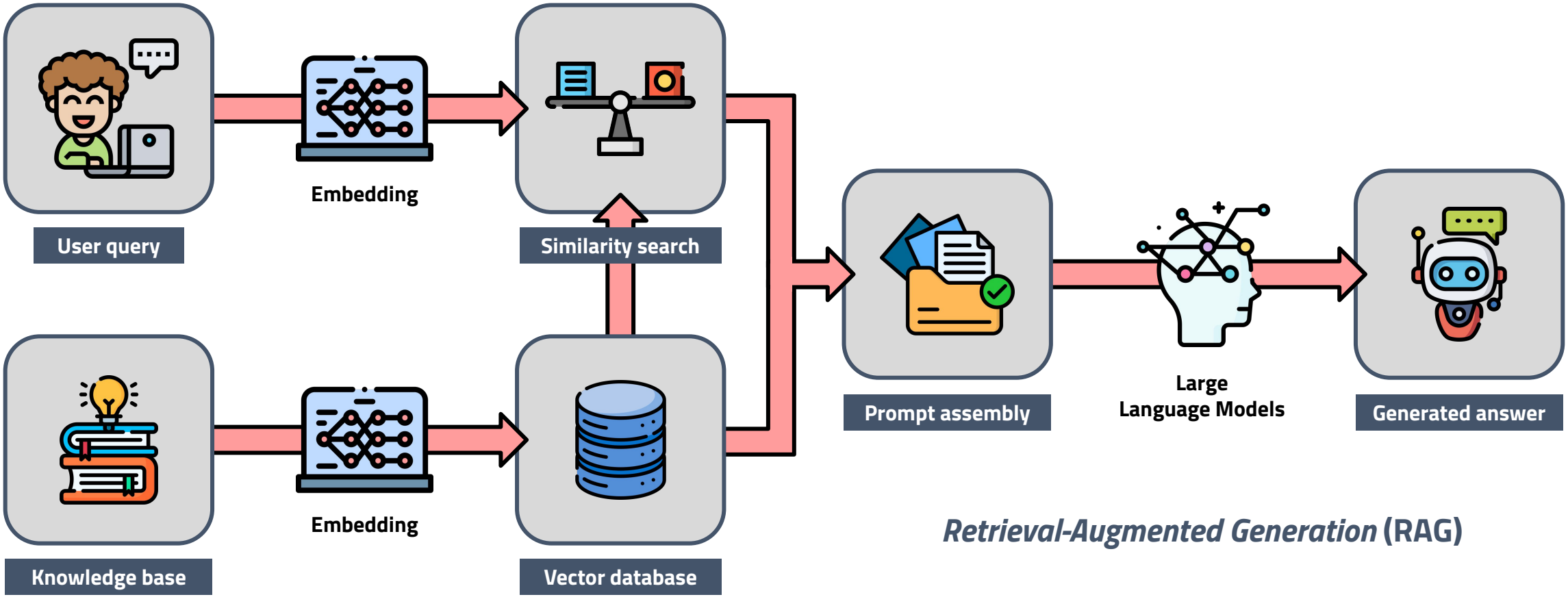
Risorse necessarie per: *labelling + data cleaning + embedding*



Risorse necessarie per: addestrare i 64 modelli di classificazione

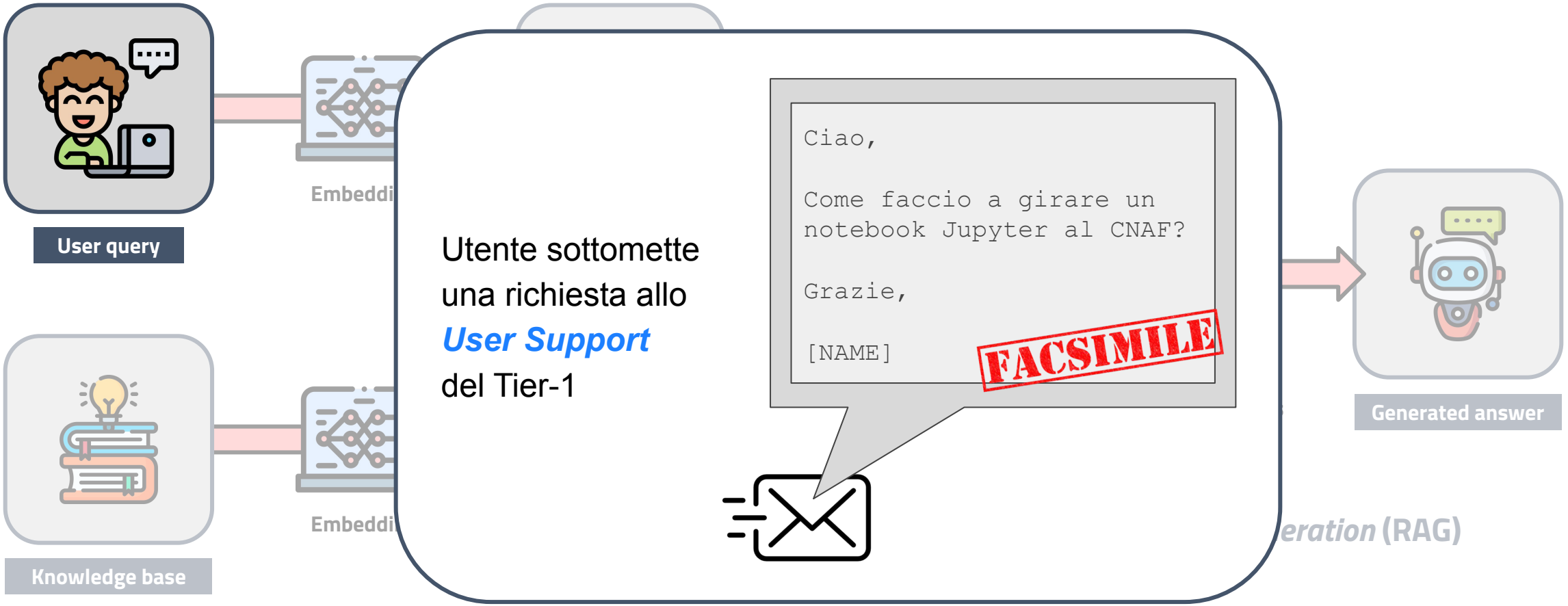
USER SUPPORTER DIGITALE

Come costruire uno User Supporter digitale



***Retrieval-Augmented Generation* (RAG)**

Come costruire uno User Supporter digitale



Come costruire uno User Supporter digitale

User query

Knowledge base

Embedding

Vecto

Tier1 - Documentation

STRUTTURA AD ALBERO DELLA PAGINA

- INFN-CNAF Tier-1 User Guide (April 2024 - v17)
  - 1 - CNAF
  - 2 - Tier-1
  - 3 - Bastion & user interfaces
  - 4 - Farming
  - 5 - Storage
  - 6 - The HPC cluster
  - 7 - Cloud @ CNAF
  - 8 - Digital Personal Certificates and Proxies ma
  - 9 - Job submission
  - 10 - Data Transfers
  - 11 - Monitoring
  - 12 - Helpful information and tips
  - 13 - Support
  - 14 - Problem report

Dashboard / Tier1 - Documentation

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:

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

n 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:

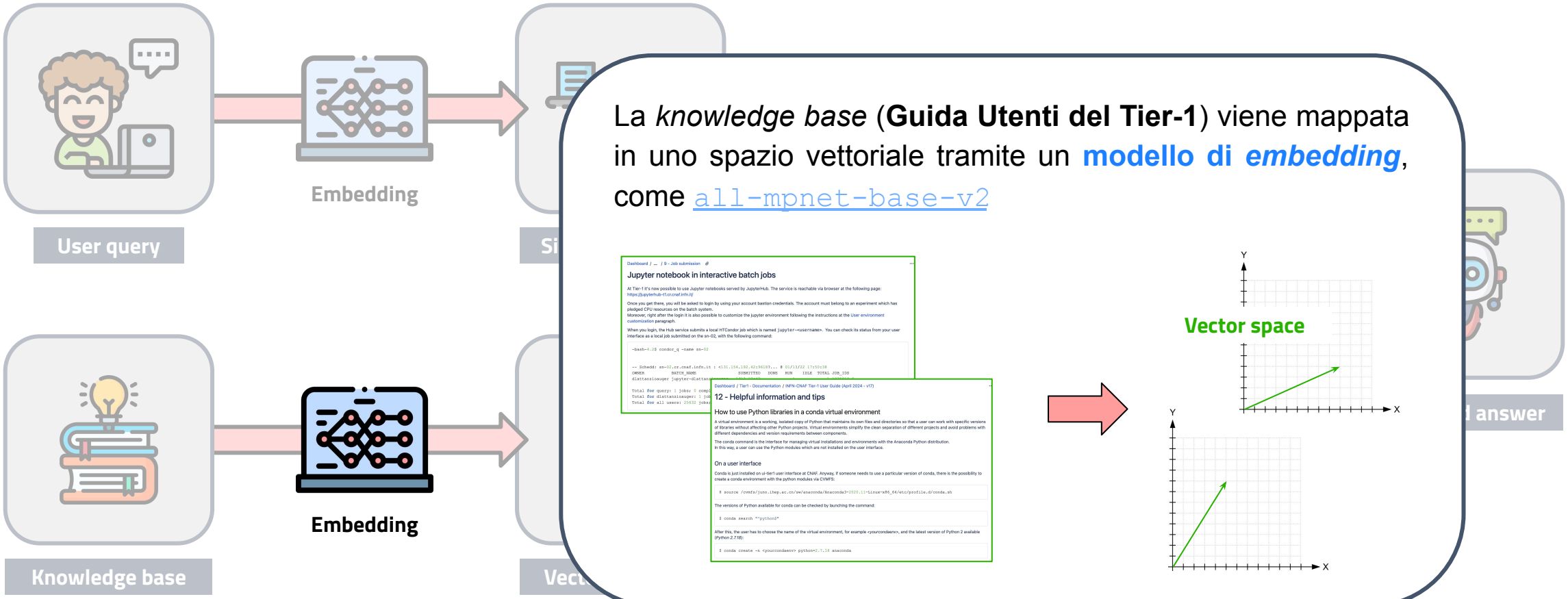
```
-bash-4.2$ condor_q -name sn-02


-- 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
```

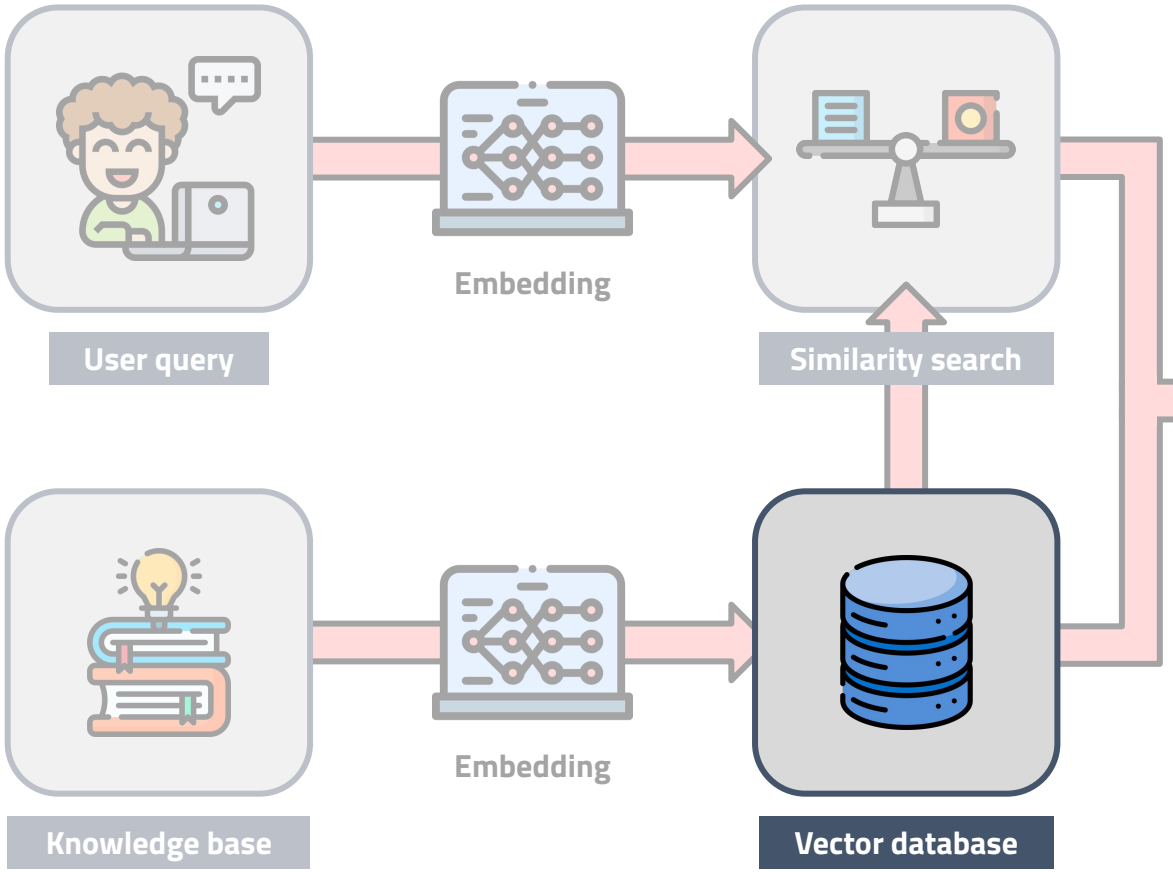
ed Generation (RAG)

Come costruire uno User Supporter digitale



La *knowledge base* (**Guida Utenti del Tier-1**) viene mappata in uno spazio vettoriale tramite un **modello di *embedding***, come all-mpnet-base-v2

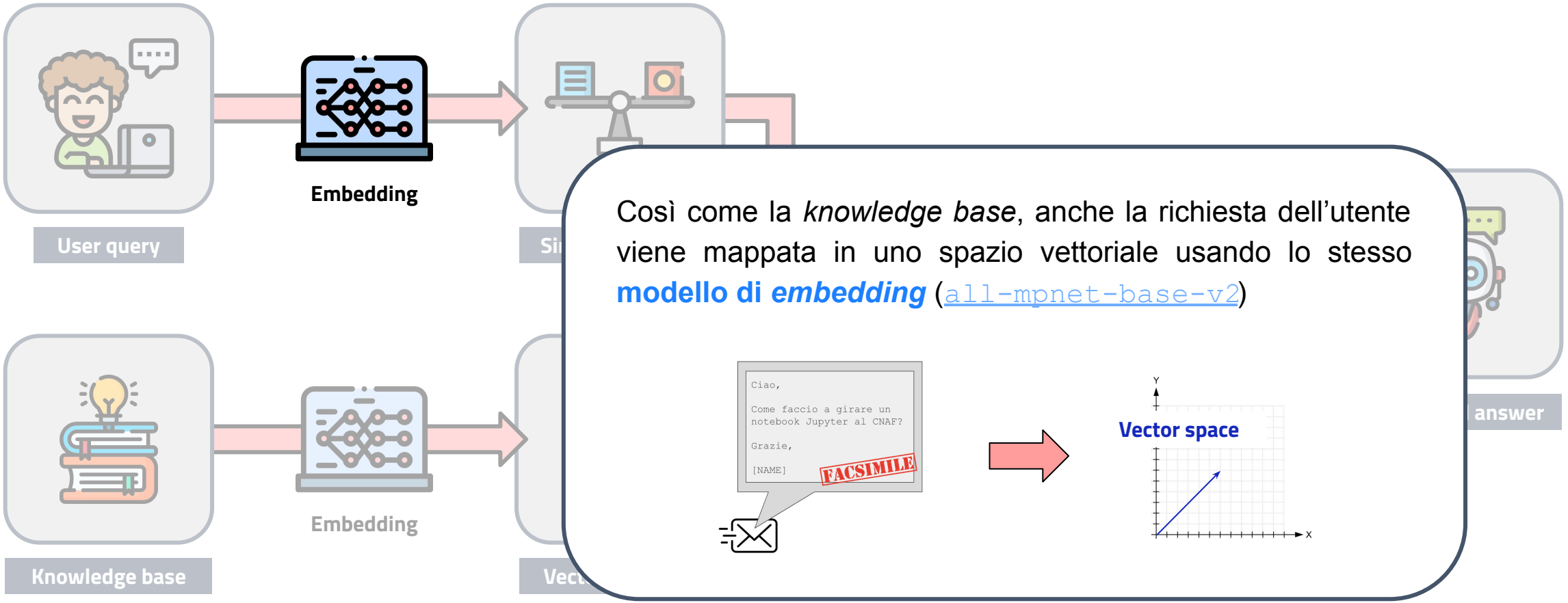
Come costruire uno User Supporter digitale



Per salvare il risultato dell'*embedding* occorre un **database** ottimizzato per **conservare vettori** di grandi dimensioni e i loro ***metadata*** (es.: testo sorgente), oltre a permettere una **ricerca efficiente** nello spazio vettoriale (caratteristiche simili sono alla base di Chroma)

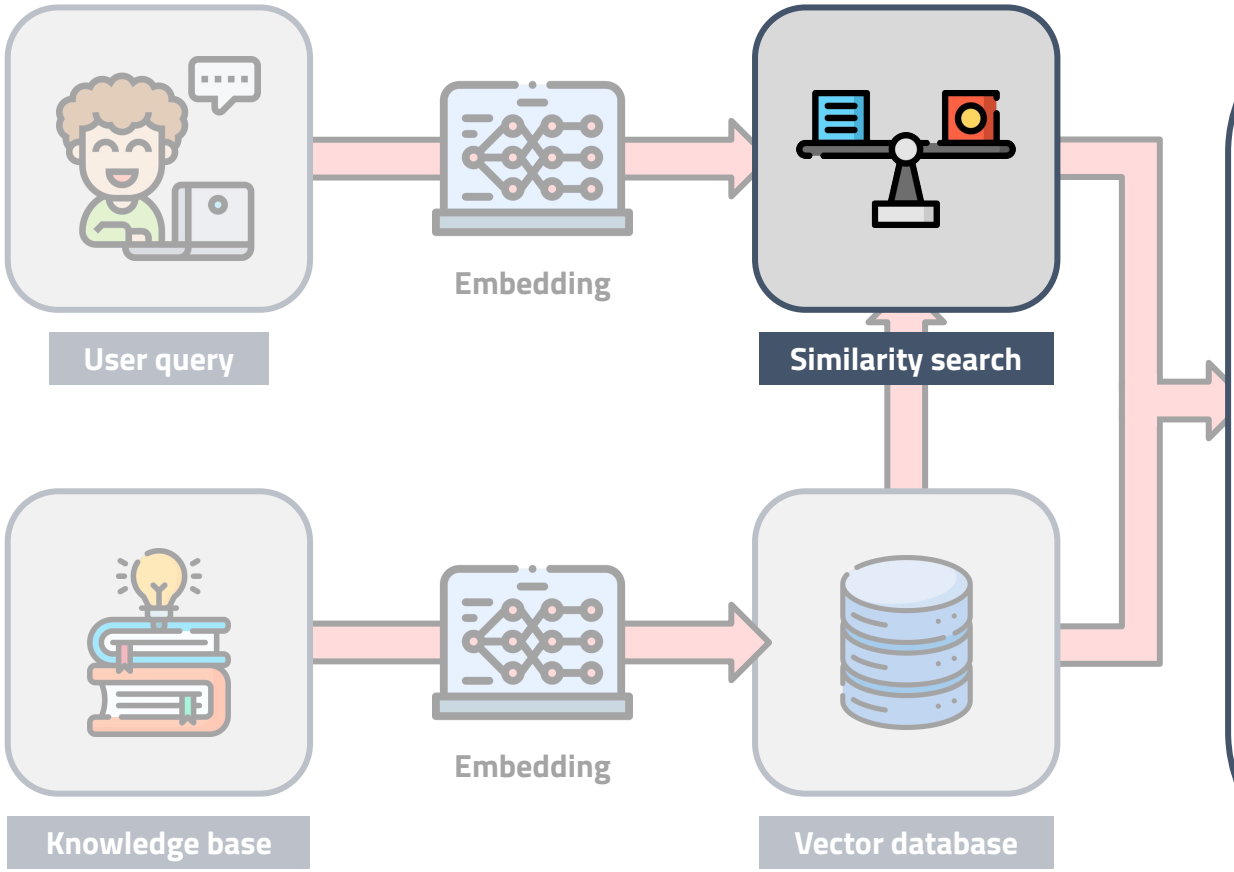
Chroma

Come costruire uno User Supporter digitale

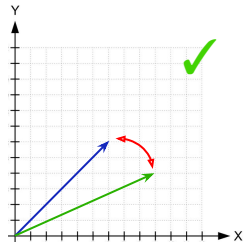


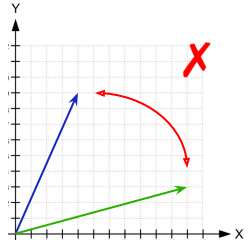
Così come la *knowledge base*, anche la richiesta dell'utente viene mappata in uno spazio vettoriale usando lo stesso **modello di *embedding*** (`all-mpnet-base-v2`)

Come costruire uno User Supporter digitale



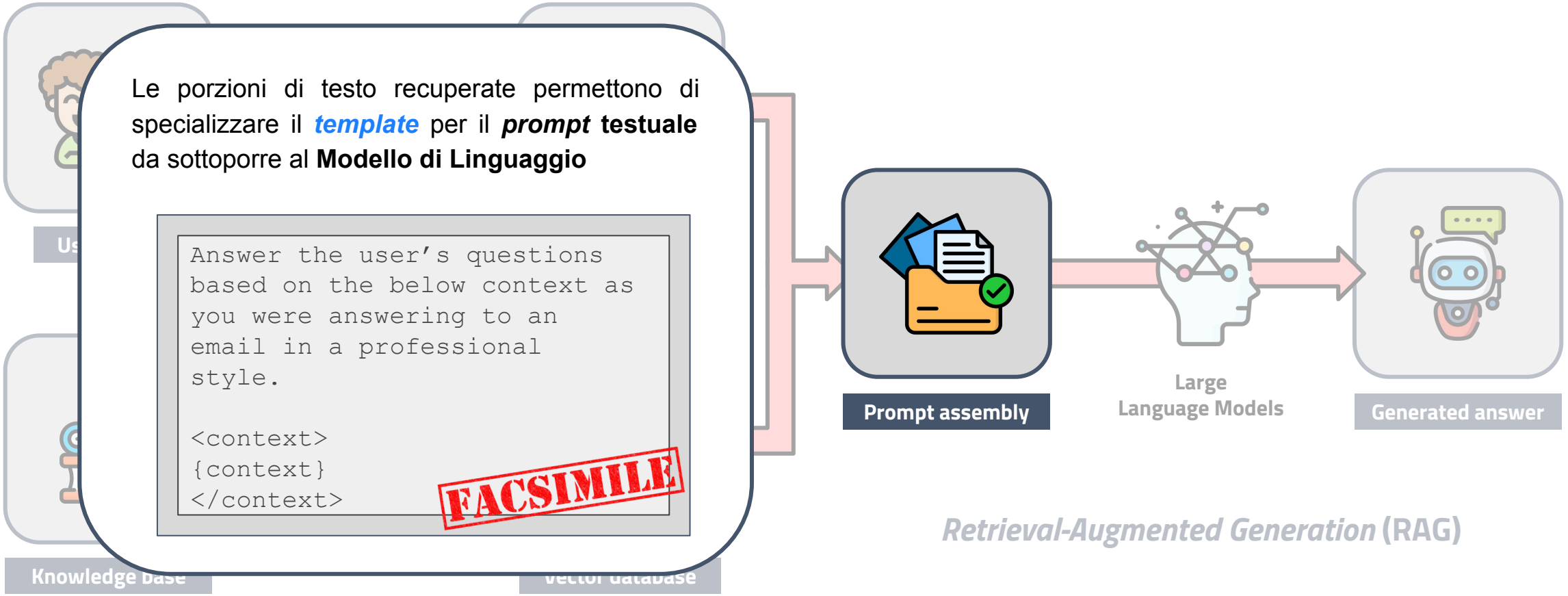
Per recuperare dalla **guida** le info pertinenti per rispondere alla **richiesta dell'utente** possiamo **confrontare per similarità** (es.: prodotto scalare) il risultato dell'*embedding* nei due casi





Trovati gli *embedding* della guida che meglio aderiscono all'*embedding* della richiesta, il **database vettoriale** può essere interrogato per ***metadata*** corrispondenti

Come costruire uno User Supporter digitale



Retrieval-Augmented Generation (RAG)

Come costruire uno User Supporter digitale

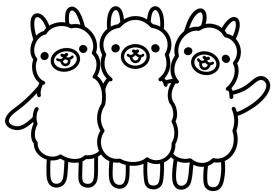
User query

Knowledge base

Il *prompt* creato può quindi essere sottomesso a un **Grande Modello di Linguaggio** (LLM) che, anche se non "esperto" di uno specifico argomento, è capace di estrarre il **contenuto "interessante"** da `{context}` e di riformularlo

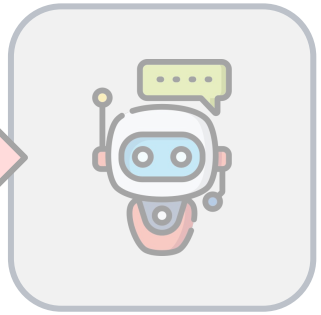
Potendo fare a meno del *fine-tuning*, gli LLM sono spesso descritti come delle tecnologie *general-purpose* e dunque chiamati anche **Modelli Fondazionali** [5]

Meta



`ollama run llama3`[2]

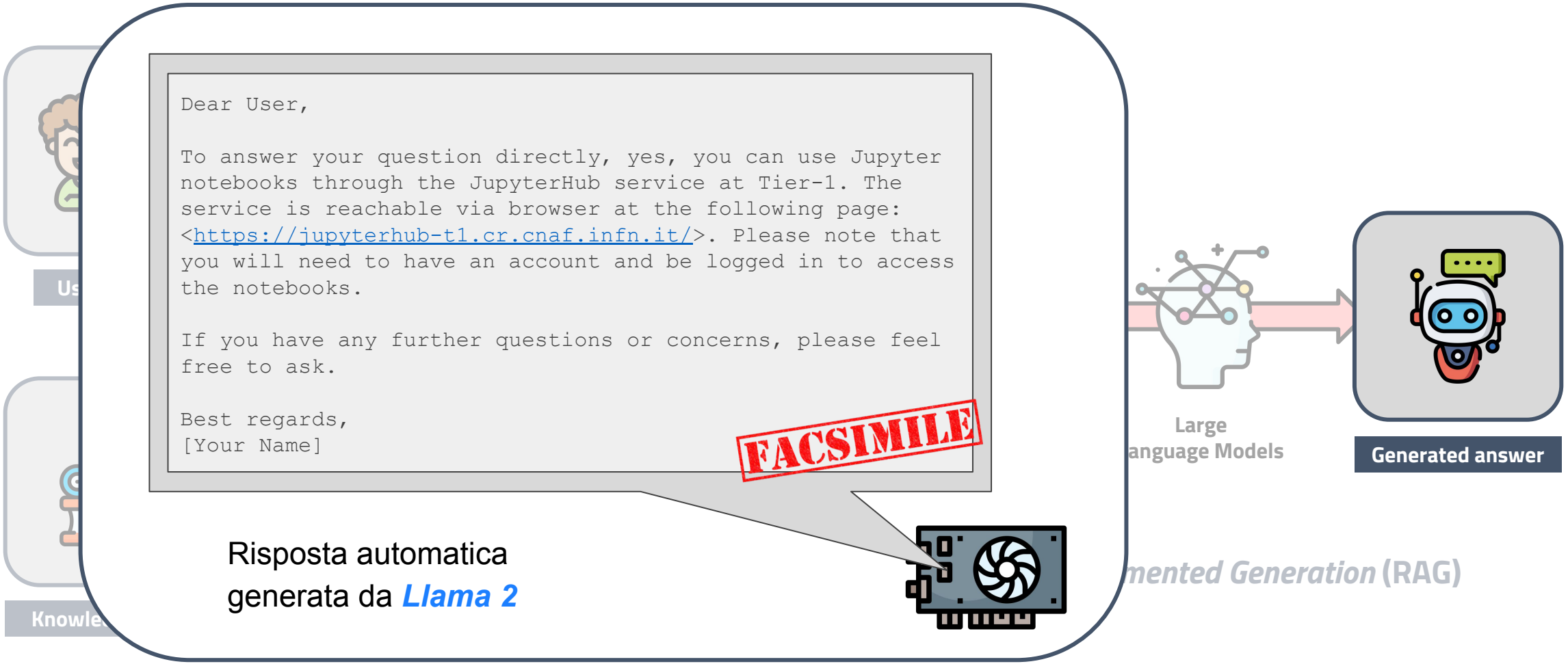
Large Language Models



Generated answer

gmented Generation (RAG)

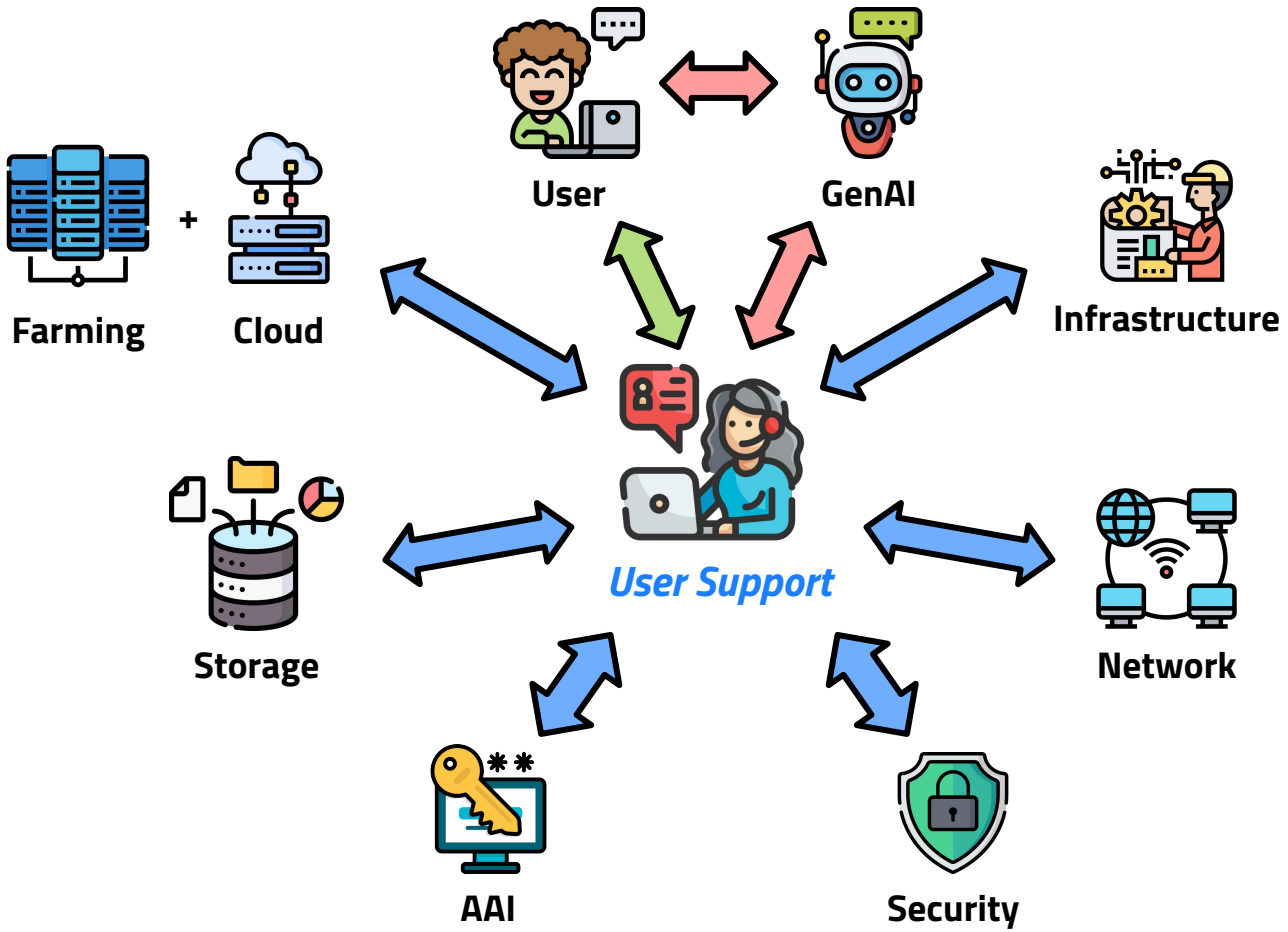
Come costruire uno User Supporter digitale



Risposta automatica generata da ***Llama 2***

CONSIDERAZIONI FINALI

Conclusioni e prossimi step



- Il crescente numero di utenti attivi al Tier-1 e l'adozione di nuove tecnologie rende il **ruolo dello *User Support*** sempre più cruciale
- Alcune soluzioni basate su **tecniche di AI** sono state investigate per evolvere lo *User Support*
- **Primi risultati incoraggianti** anche se c'è spazio per migliorare algoritmi/modelli
  - Modello RAG a **"multi-esperti"** specializzati sui problemi del singolo reparto (es.: *template custom* per *prompt* testuale, *fine-tuning* degli LLM)
  - Risultato della *similarity search* impiegabile per **"valutare" la *knowledge base*** e, eventualmente, "suggerire" migliorie/aggiornamenti

Grazie per l'attenzione!

Ci sono domande, commenti o suggerimenti?

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

Elisabetta Ronchieri (INFN CNAF)
email: elisabetta.ronchieri@cnaf.infn.it

Alberto Trashaj (Università di Bologna)
email: alberto.trashaj@studio.unibo.it

Carmelo Pellegrino (INFN CNAF)
email: carmelo.pellegrino@cnaf.infn.it

Bibliografia e sitografia

1. INFN-CNAF website, https://www.cnaf.infn.it
2. D. Cesini *et al.*, “Migrating the INFN-CNAF datacenter to the Bologna Tecnopolo: a status update”, in *26th International Conference on Computing in High Energy & Nuclear Physics (CHEP 2023)*
3. C. Pellegrino *et al.*, “Support for experiments at INFN-T1”, in *26th International Conference on Computing in High Energy & Nuclear Physics (CHEP 2023)*
4. E. Ronchieri *et al.*, “An Artificial Intelligence-based service to automatize the INFN CNAF User Support”, in International Symposium on Grids & Clouds (ISGC) 2024
5. R. Bommasani *et al.*, “On the Opportunities and Risks of Foundation Models”, arXiv:2108.07258

BACKUP

Stagionalità delle email allo User Support