






A CERN Knowledge Retrieval Chatbot

(Harnessing the power of AI for efficient information retrieval)

Verena Kain, Florian Rehm, Sofia Vallecorsa, Giovanni Guerrieri and **Manuel Guijarro**

Introduction

- Problem:
 - Finding relevant information within CERN's extensive internal documentation is complex and time-consuming.
- Solution:
 - AccGPT  - an AI-powered chatbot that leverages Natural Language Processing (NLP) for knowledge retrieval.
- Goal:
 - Create a purpose-built chatbot using open-source Large Language Models (LLMs).
- Future Potential:
 - Assist in code development and other tasks.

Motivation and Demand

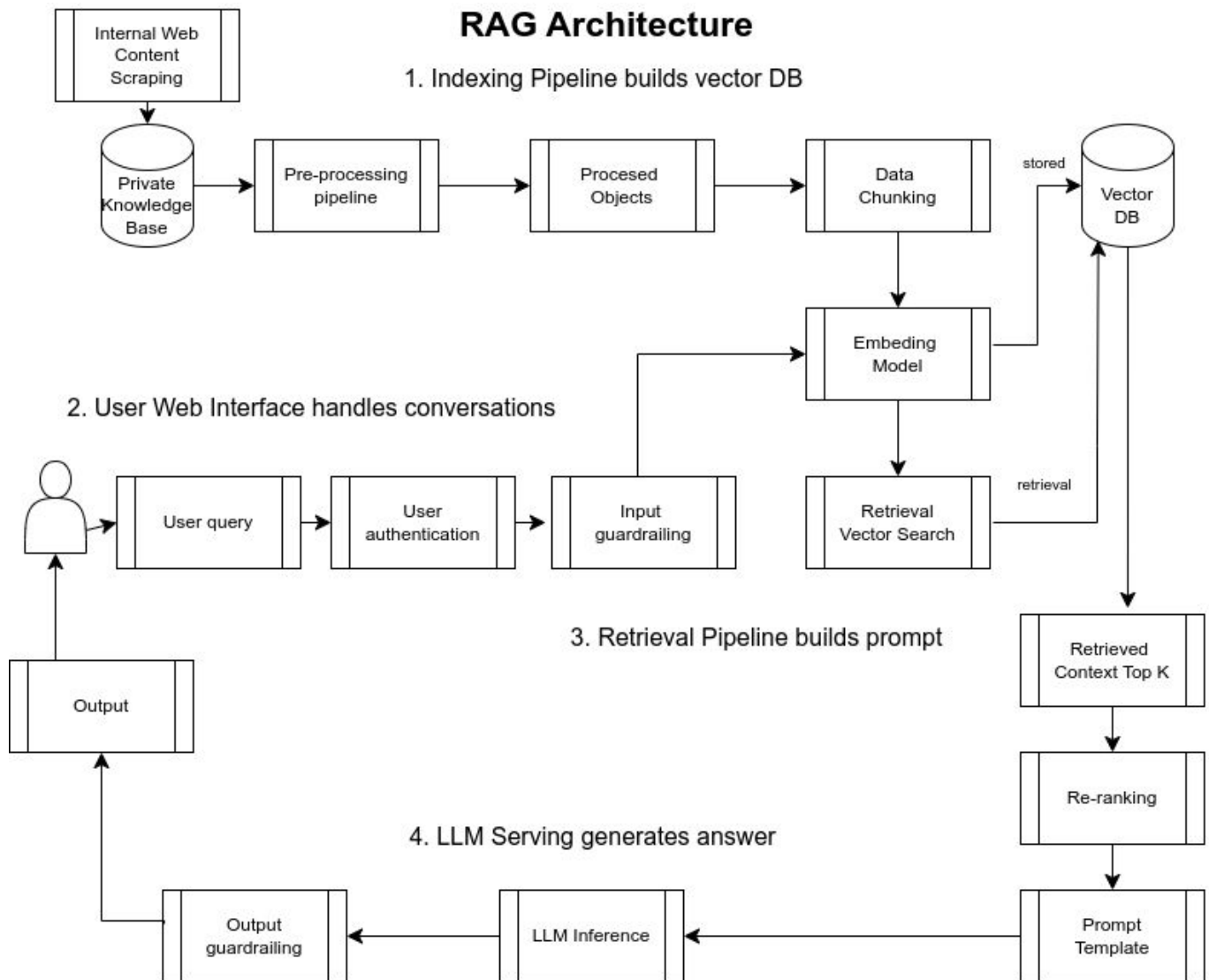
- Growing demand for LLMs at CERN for various applications:
 - AI Chatbots
 - Information Retrieval: User Support, HR selection processes, ...
 - Summarization and Minutes generation
 - Code Assistants (IDE integrated)
 - Code Generation
 - SQL Generator/Executor
 - Application use cases:
 - Zenodo Information Categorization
 - Transcription and Translation Services (TTaaS)
- Easy and efficient access to LLMs is crucial for widespread adoption of GenAI

Prototype and Architecture

➤ Prototype demonstrates successful knowledge retrieval using a RAG pipeline

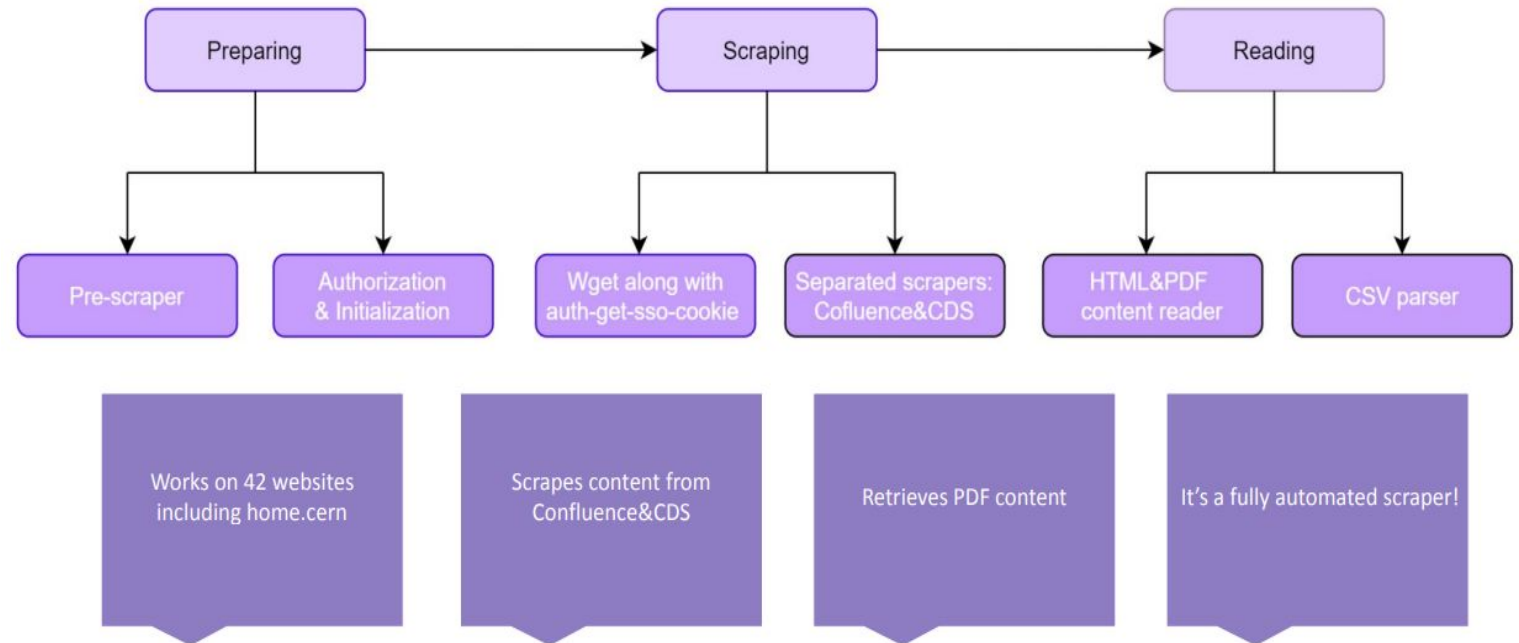
➤ Key Components of RAG:

- Indexing Pipeline
- Retrieval Pipeline
- LLM serving
- User Web Interface



Indexing Pipeline: Functions

- **Web Content Scraping:** Collects data from CERN's internal web content, including IT service documentation and parts of the public CERN website.
- **Preprocessing:** Cleans data, removes irrelevant information, and structures it in a consistent format (JSON).
- **Chunking:** Divides data into smaller, manageable units for the LLM.
- **Vectorization:** Embeds text chunks into vectors to represent their meaning.



Scraping and Sources

➤ Data Sources:

- Internal web pages (HTML, PDF)
- Confluence spaces (using a specialized scraper with Python Selenium and API interaction)
- CERN Document Server (CDS) - scientific articles (CDS XML API)

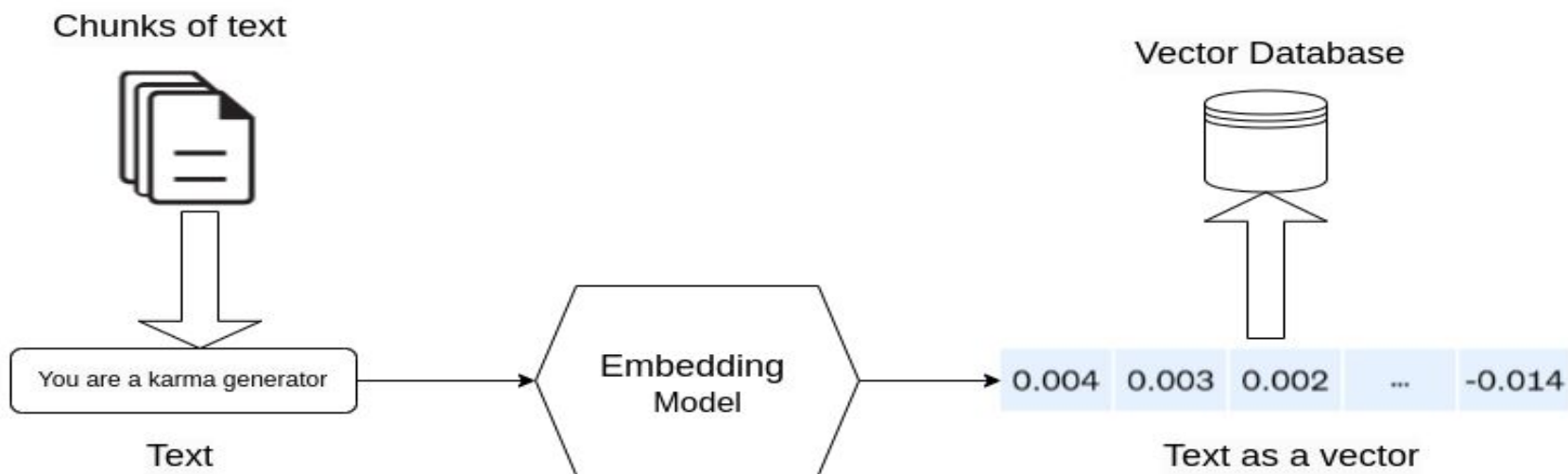
➤ Future:

- Expansion to include more internal CERN domains and data formats
- Multimodal: Images, audio, video

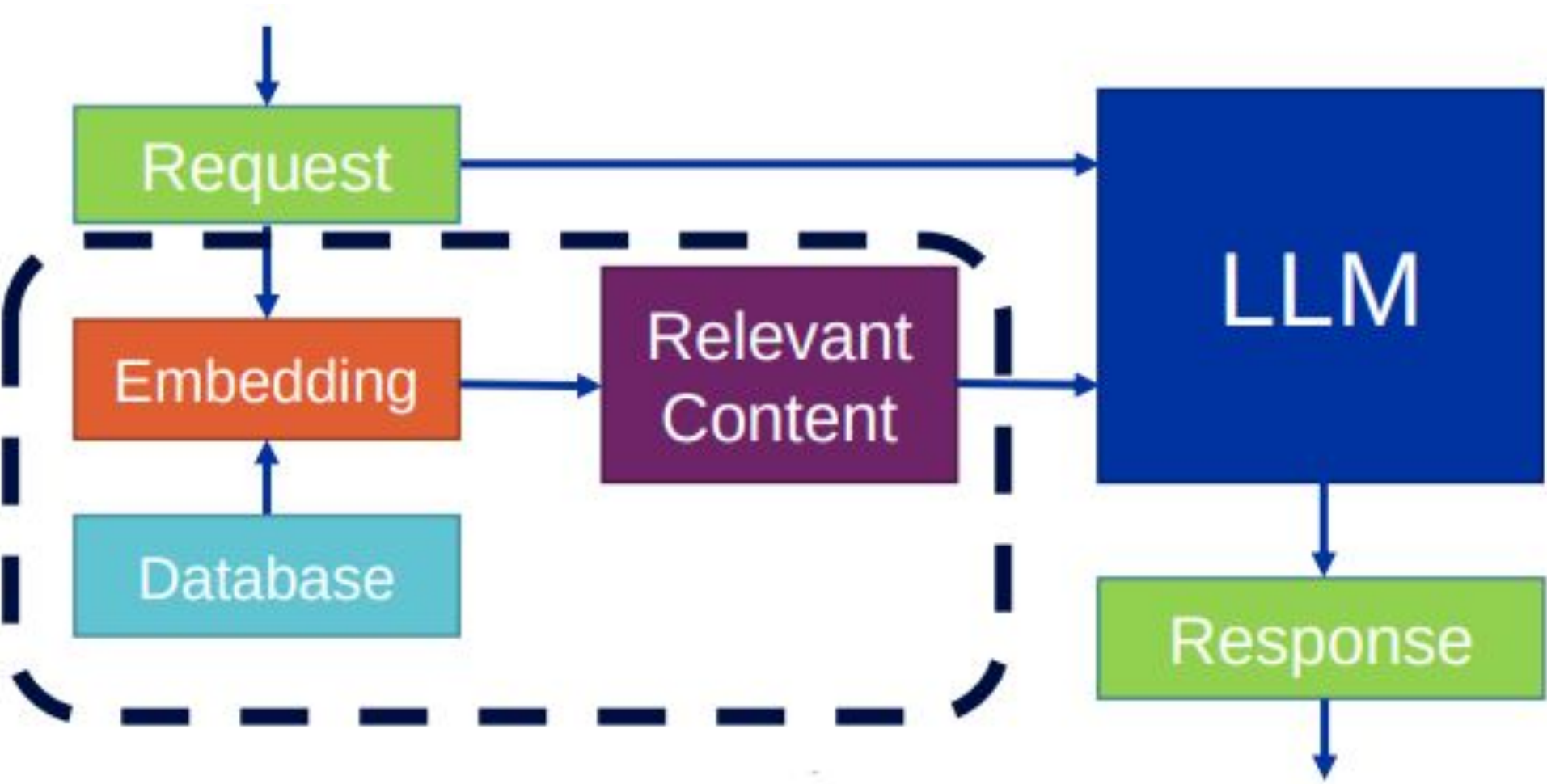
Name of column	Content of column
Index	Individual page numbers increasing by 1 (good practise)
Name	Value of tag <title>, which is unique for each web page or title of PDF file
Text	Retrieved content for web page or PDF file
Url	URL address of web page or PDF file

Chunking and Embeddings

- Chunking is crucial due to the limited input context length of LLMs.
- Current method: Rule-based division by paragraphs or sentences, with plans to explore advanced techniques.
- Embedding model: **e5-large-v2**, chosen for its efficiency, effectiveness, and multilingual capabilities.
- Vector database: **ChromaDB**, used for storing and retrieving embeddings and associated metadata.



Core of Retrieval Pipeline

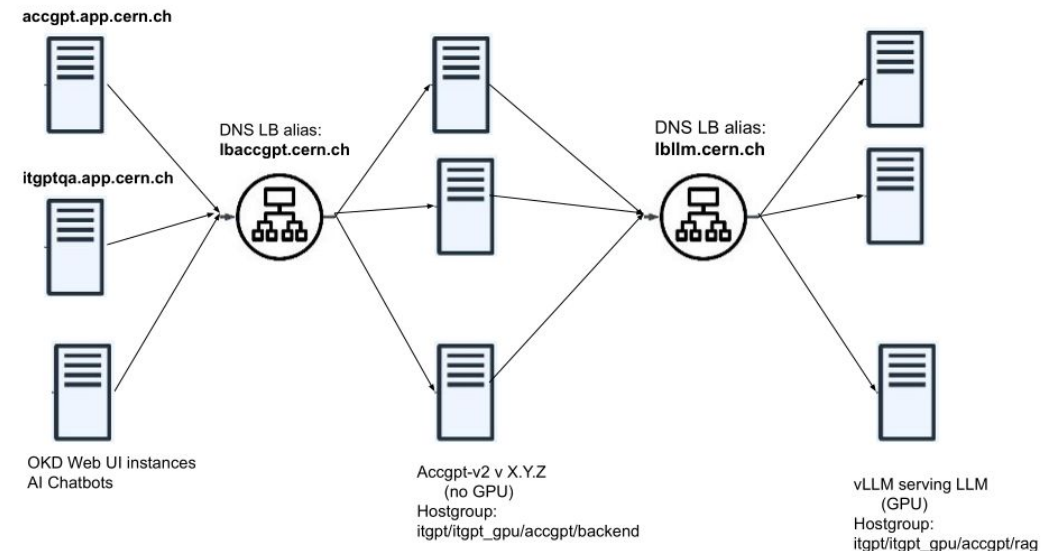
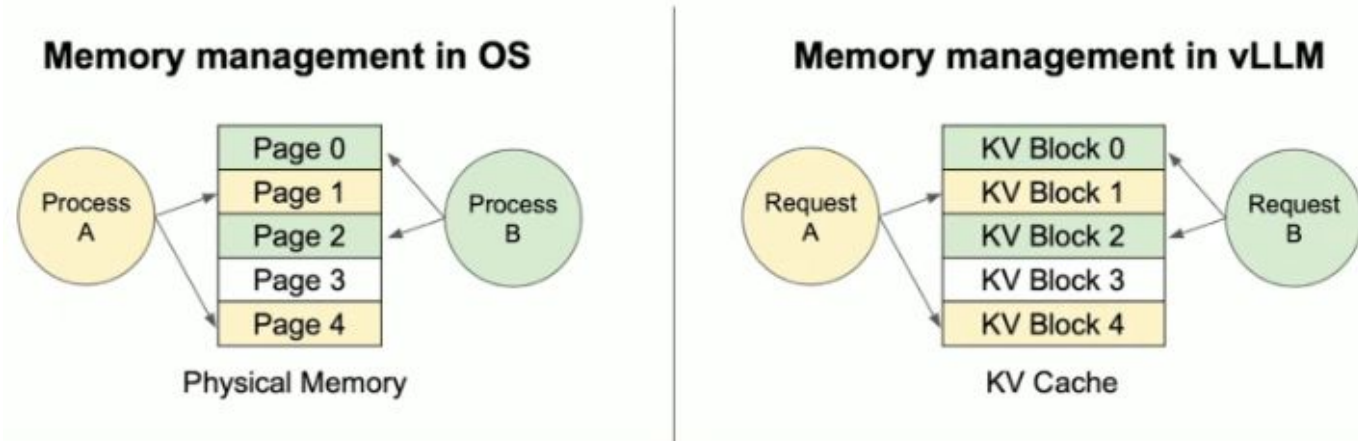


Retrieval Pipeline

- **Purpose:** Retrieve relevant knowledge from the vector database to support the LLM in answering user questions.
- **Process:**
 - Converts the user's question into a vector using the same embedding model as used for creating the embedding database.
 - Performs a similarity search between the question vector and vectors in the database.
 - Retrieves the most relevant text chunks.
- **Enhancements for Accuracy:**
 - Additional embedding search using keywords extracted from the question by an LLM.
 - Implementation of a re-ranker model for more precise filtering and confidence scores.

LLM Serving

- **Functionality:** Provides an API for interacting with the LLM, sending prompts and receiving responses.
- **Implementation:** vLLM library, featuring PagedAttention for efficient memory management.
- **Benefits of vLLM:**
 - Reduced memory overhead
 - High throughput and performance
 - Compatibility with Hugging Face models
 - Support for multi-modal workloads
- **Current LLM:** Meta-Llama-3.1-8B-Instruct, chosen due to limitations in available GPU resources.

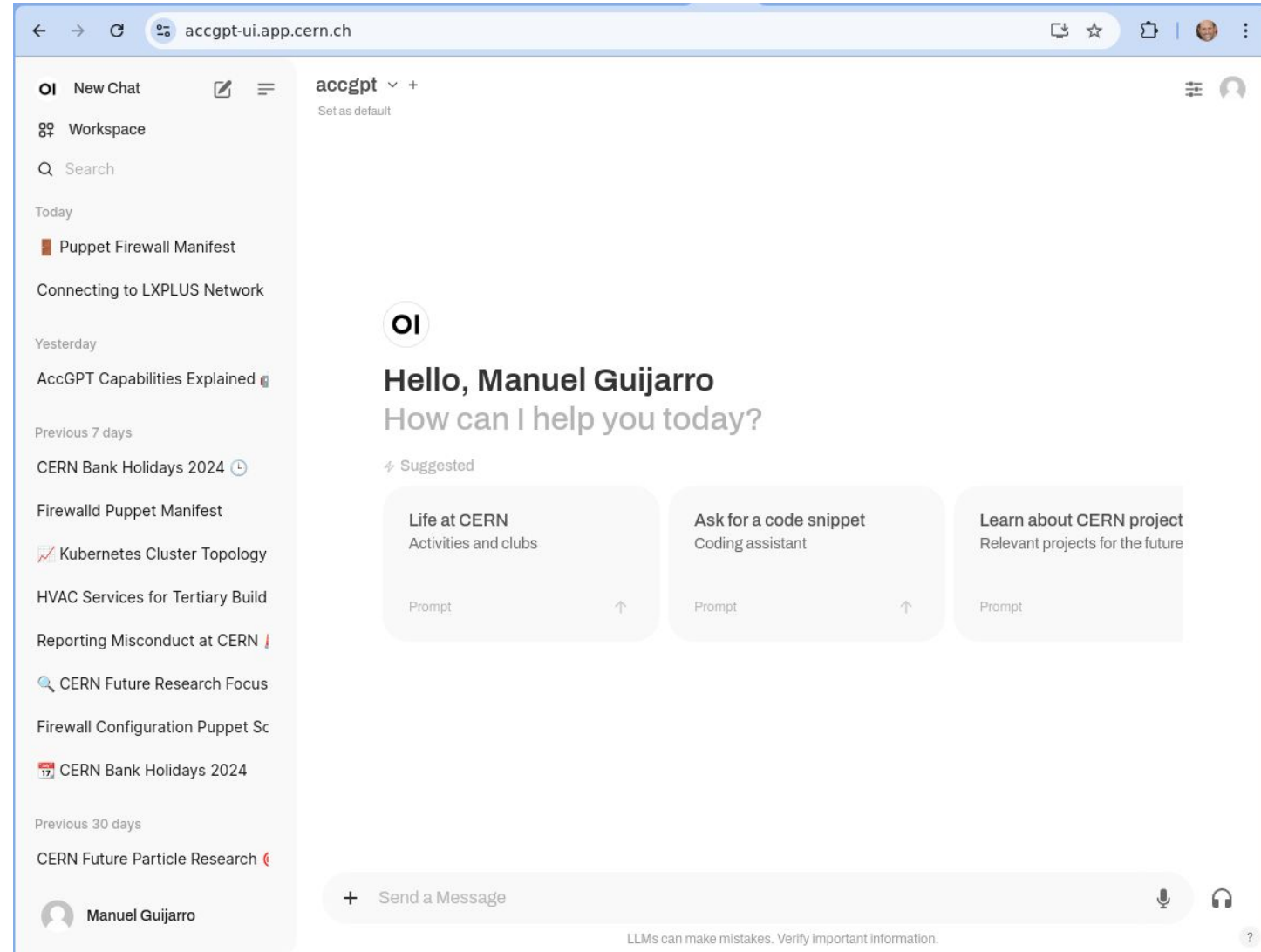


Challenges with Commercial LLMs

- CERN's Cloud Policy: Compliance with data privacy and security requirements poses challenges.
- User Preferences: CERN users prefer on-site, self-hosted models for continuous operation, confidentiality, and control over data.
- Fine-tuning Concerns: Reluctance to fine-tune commercial models due to ownership and exit strategy issues.
- Cost Considerations:
 - Evaluating costs, performance, and reliability is key when choosing between self-hosting LLMs or commercial providers like OpenAI.
 - For high-volume usage, self-hosted LLMs can become more cost-effective:
 - 720 questions = 20,105 input tokens + 1,170 output tokens
 - Cost: 1 hour of an A100 GPU vs \$8.5 OpenAI subscription

User Web Interface

- **Platform:** OpenWebUI, an open-source solution chosen for its customizability, features, and community support.
- **Benefits of OpenWebUI:**
 - Integration with commercial and self-hosted LLMs
 - Cross-device compatibility
 - Scalability
- **Deployment:** OpenShift instance for scalability, security, and efficiency.



Screenshot: DBOD example

The screenshot shows a web browser window with the URL `accgpt-ui.app.cern.ch/c/4579634d-767c-471f-9df1-8b2819aaf0f5`. The chat interface includes a sidebar with options like 'New Chat', 'Workspace', and 'Search'. The main chat area shows a user asking 'How can I upgrade my DBOD instance?' and an AI response from 'accgpt' providing a list of steps for upgrading a DBOD instance. A 'Citation' popup is overlaid on the bottom of the chat, displaying the source URL and a confidence score of 81%.

accgpt-ui.app.cern.ch/c/4579634d-767c-471f-9df1-8b2819aaf0f5

New Chat Workspace Search

Today

DBOD Upgrade Guidance

OpenStack Flavor Modification

VM Capacity Upgrade

Circulaire de conduire CERN

Yesterday

AIADM Machine Status Check

CERN Ba

CERNs F

Pupp

OC11

How can I upgrade my DBOD instance?

accgpt

To upgrade your DBOD instance, follow these steps:

1. Validate the upgrade process in a test instance or by creating a clone of your current production instance.
2. Run the upgrade checker report to identify and correct any errors.
3. If the upgrade checker report does not contain any errors, you can run the upgrade.
4. If the upgrade option is still greyed out, contact DBOD support.

Note: It is recommended to follow the recommended upgrade workflow, which includes

Citation

Source
https://dbod-user-guide.web.cern.ch/instance_management/upgrading_your_instance_from_the_web_portal/index.html

Content

Confidence: 81%

Screenshot: Add context !!!

The screenshot shows a web browser window with the URL `accgpt-ui.app.cern.ch/c/9066ba14-36f8-46a4-a58b-d5693756aaaa`. The interface is a chat application. On the left is a sidebar with a search bar and a list of chat items, including "CERN Bank Holidays 2024" which is selected. The main chat area shows a user input: "What are official bank holidays at CERN in 2024?". The AI response, from the model "accgpt", lists the following holidays:

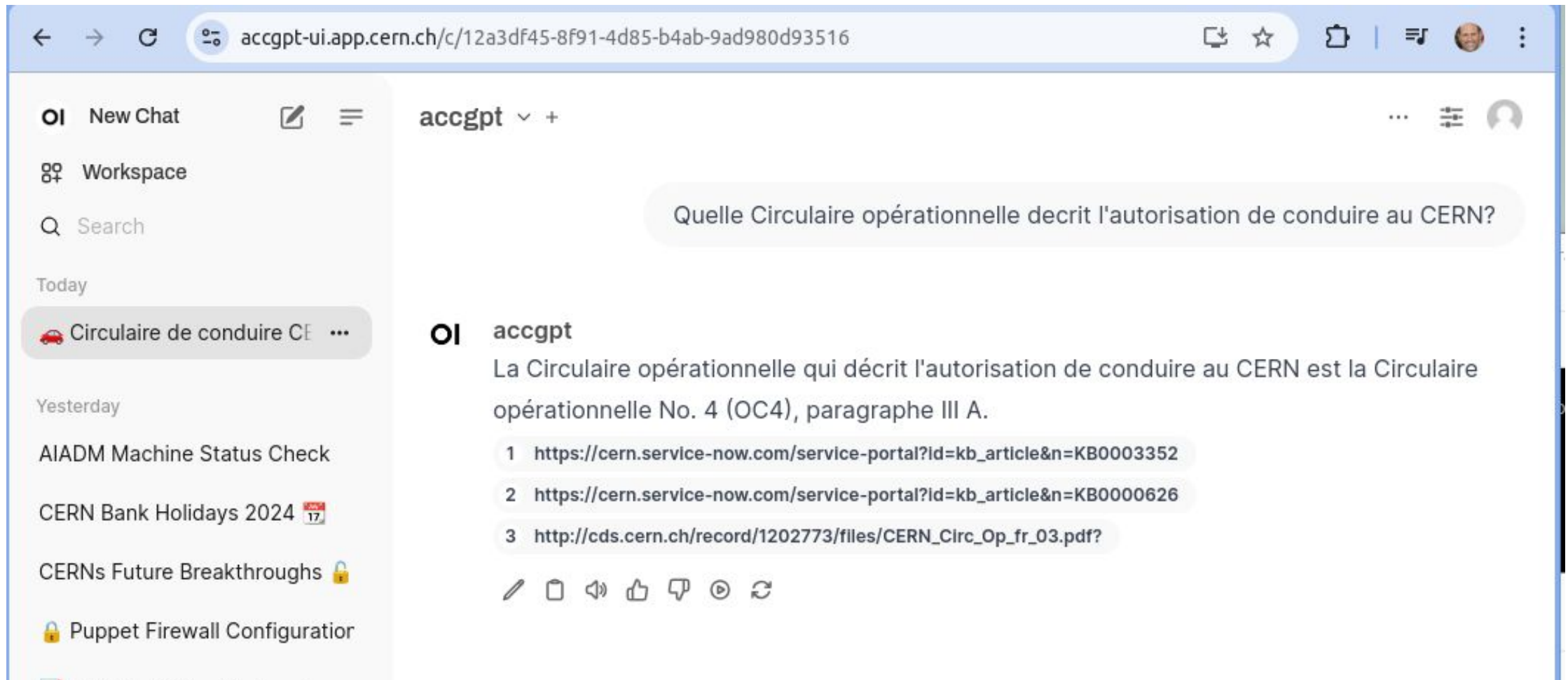
- Monday, 1 January (New Year)
- Friday, 29 March (Good Friday)
- Monday, 1 April (Easter Monday)
- Wednesday, 1 May (1 May)
- Thursday, 9 May (Ascension day)
- Monday, 20 May (Whit Monday)
- Thursday, 5 September ("Jeûne genevois")
- Tuesday, 24 December (Christmas Eve)
- Wednesday, 25 December (Christmas)
- Tuesday, 31 December (New Year's Eve)

The response also includes a note: "Additionally, the Laboratory will be closed from Saturday, 21 December 2024 until Sunday, 5 January 2025 inclusive, during the end-of-year holiday period." A link is provided: <https://home.cern/news/official-news/cern/official-holidays-2024-and-end-year-closure-20242025>. At the bottom, there is a "Send a Message" input field and a disclaimer: "LLMs can make mistakes. Verify important information."

Screenshot: No real French Support

The screenshot shows a web browser window with the URL `accgpt-ui.app.cern.ch/c/a6cb6aa0-65e3-456a-a55d-055c7e98a187`. The interface is a chat application with a sidebar on the left containing a list of chat topics such as 'Hadoop Data Ingestion', 'OpenSearch Index Config', and 'DBOD Upgrade Guidance'. The main chat area shows a conversation with a user named 'accgpt'. The user asks: "How can I ingest data into hadoop?". The AI assistant responds in English, recommending frameworks like Apache Spark, Apache Flume, and Apache Kafka Connect, and providing a link to a CERN document: `https://hadoop.docs.cern.ch/hdfs/ingesting_the_data/index.html`. Below this, the user asks in French: "Comment puis-je ingérer des données dans Hadoop?". The AI assistant responds in French, stating that no relevant documents were found and suggesting the user rephrase their question or contact developers.

Screenshot: ...only if French doc



← → ↻ accgpt-ui.app.cern.ch/c/12a3df45-8f91-4d85-b4ab-9ad980d93516 📄 ☆ 🗂️ | 📖 👤 ⋮

OI New Chat ✍️ ☰ **accgpt** ▾ + ⋮ ☰ 👤

🗂️ Workspace

🔍 Search

Today

🚗 Circulaire de conduire CE ...

Yesterday

AIADM Machine Status Check

CERN Bank Holidays 2024 📅 17

CERNs Future Breakthroughs 🔒

🔒 Puppet Firewall Configurati...

Quelle Circulaire opérationnelle décrit l'autorisation de conduire au CERN?

OI **accgpt**

La Circulaire opérationnelle qui décrit l'autorisation de conduire au CERN est la Circulaire opérationnelle No. 4 (OC4), paragraphe III A.

- 1 https://cern.service-now.com/service-portal?id=kb_article&n=KB0003352
- 2 https://cern.service-now.com/service-portal?id=kb_article&n=KB0000626
- 3 http://cds.cern.ch/record/1202773/files/CERN_Circ_Op_fr_03.pdf?

✍️ 📄 🔊 👍 🗨️ 🎥 🔄

Screenshot: OpenStack example

The screenshot shows a web browser window with the URL `accgpt-ui.app.cern.ch/c/39ee9973-d39f-402a-9868-bd1310386ba0`. The interface is a chat application with a sidebar on the left and a main chat area on the right.

Chat History (Left Sidebar):

- New Chat
- Workspace
- Search
- Today
 - OpenStack Flavor Modificati ...
 - VM Capacity Upgrade
 - Circulaire de conduire CERN
- Yesterday
 - AIADM Machine Status Check
 - CERN Bank Holidays 2024
 - CERNs Future Breakthroughs
 - Puppet Firewall Configurior
 - OC11 Definition Explanation
- Previous 7 days
 - CERN Secret Management Guide
 - Installing Mac at CERN

Current Chat:

User: I need more resources for my virtual machine, how do I change the flavor in OpenStack?

AI (accgpt): To change the flavor in OpenStack, you can use the `openstack flavor list` command to view the available flavors and their IDs. Then, you can use the `openstack server update` command to update the instance with the new flavor ID. For example:

```
openstack flavor list
openstack server update <instance_id> --flavor <new_flavor_id>
```

Note that you need to replace `<instance_id>` with the actual ID of the instance you want to update, and `<new_flavor_id>` with the ID of the new flavor you want to use.

Also, keep in mind that additional flavors are available upon request and approval, and you can open a ticket to the Cloud team to request a new flavor.

1 https://clouddocs.web.cern.ch/using_openstack/resizing_a_vm.html

Screenshot: CERNBox example

The screenshot shows a web browser window with the URL `accgpt-ui.app.cern.ch/c/63ea3a8d-9e82-4c25-b79f-18ff485a8c49`. The interface is a chat application for 'accgpt'. On the left is a sidebar with a list of items: 'New Chat', 'Workspace', 'Search', and a list of documents including 'CERNBox Access Manag...', 'Hadoop Data Ingestion Guide', 'OpenSearch Index Config', 'DBOD Upgrade Guidance', 'OpenStack Flavor Modification C', 'VM Capacity Upgrade', 'Circulaire de conduire CERN', and 'Puppet Firewall Configurator'. The main chat area shows a user asking 'How do I manage access to a project space in CERNBox?'. The AI assistant 'accgpt' responds with the following text:

To manage access to a project space in CERNBox, follow these steps:

1. Control access to the entire project via the `cernbox-project-<projectname>-writers` and `-readers` e-groups. You can check their membership at <https://egroups.cern.ch>.
2. To add or remove members from these e-groups, log in to <https://e-groups.cern.ch> using your primary account.
3. To share individual folders and their sub-folders, use the Authenticated Share method in CERNBox. Only members of the `-admins` e-group can share folders.
4. To check which e-groups and users have access to the project space, run the command `eos root://eosproject.cern.ch attr ls /eos/project/<initial>/<project-name>/<path-to-folder>` on LXPLUS.

Note: The process of adding members to the project access e-groups may take a relatively long time (hours or days) due to synchronization processes upstream of CERNBox.

- 1 <https://cernbox.docs.cern.ch/web/projects/access-to-project-space/index.html>
- 2 https://cernbox.docs.cern.ch/web/projects/manage_project_space/index.html
- 3 https://cern.service-now.com/service-portal?id=kb_article&n=KB0006678

At the bottom of the chat area are icons for editing, deleting, voice, thumbs up, thumbs down, and refresh.

Screenshot: DBOD example

The screenshot shows a web browser window with the URL `accgpt-ui.app.cern.ch/c/a399f003-95b1-4c7e-8fa0-2df3384efc75`. The interface includes a sidebar with a search bar and a list of recent chat topics. The main chat area contains a user query: "How do I predefine settings and mappings for a specific index pattern in OpenSearch?". The AI response, from the user "accgpt", explains that the `index template creation` API is used for this purpose and provides an example of a REST API call to create an index template for the pattern `logs-*`. The example code is as follows:

```
bash
PUT _index_template/daily_logs
{
  "index_patterns": [
    "logs-*"
  ],
  "template": {
    "settings": {
      "number_of_shards": 2,
      "number_of_replicas": 1
    },
    "mappings": {
      "properties": {
        "timestamp": {
          "type": "date",
          "format": "yyyy-MM-dd HH:mm:ss||yyyy-MM-dd||epoch_millis"
        },
        "value": {
          "type": "double"
        }
      }
    },
    "priority": 1,
    "composed_of": [ "default_settings" ]
  }
}
```

The AI then explains that this will create an index template named `daily_logs` that applies to index patterns matching the pattern `logs-*`. It also notes that the template defines settings and mappings for the index, including the number of shards and replicas, and the data type for the `timestamp` and `value` fields. Finally, it provides instructions on how to use the `GET _index_template` API to list all index templates, or the `GET _index_template/daily_logs` API to retrieve the configuration for the specific index template.

1 https://opensearch.docs.cern.ch/cluster_admin/index_templates/index.html

Screenshot: OpenSearch example

The screenshot shows a web browser window with the URL `accgpt-ui.app.cern.ch/c/a399f003-95b1-4c7e-8fa0-2df3384efc75`. The interface includes a sidebar with a search bar and a list of chat topics. The main chat area contains a user question: "How do I predefine settings and mappings for a specific index pattern in OpenSearch?" and an AI response from "accgpt".

The AI response states: "To predefine settings and mappings for a specific index pattern in OpenSearch, you can use the `index template creation` API. Here is an example of how to create an index template for a specific index pattern:"

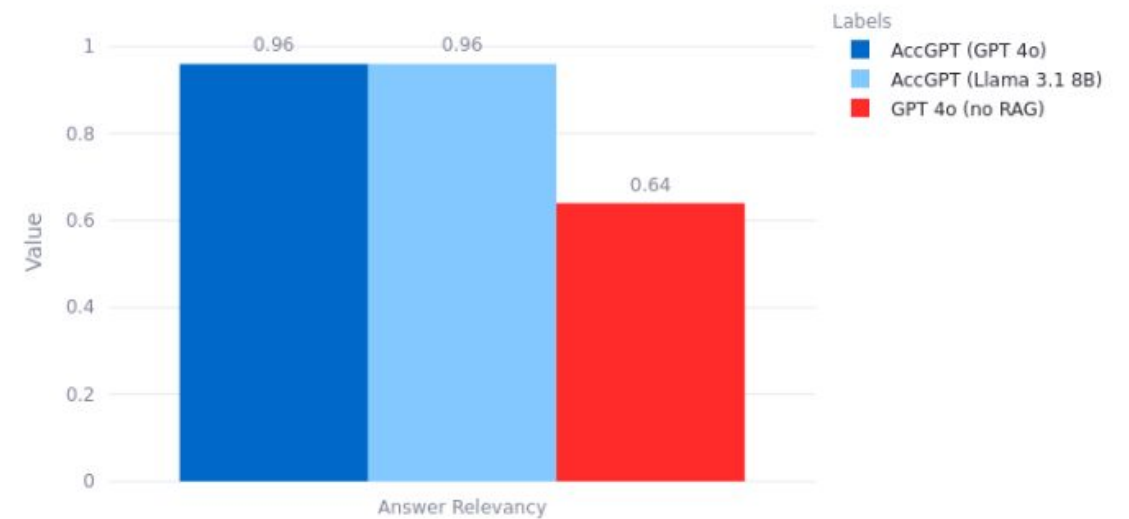
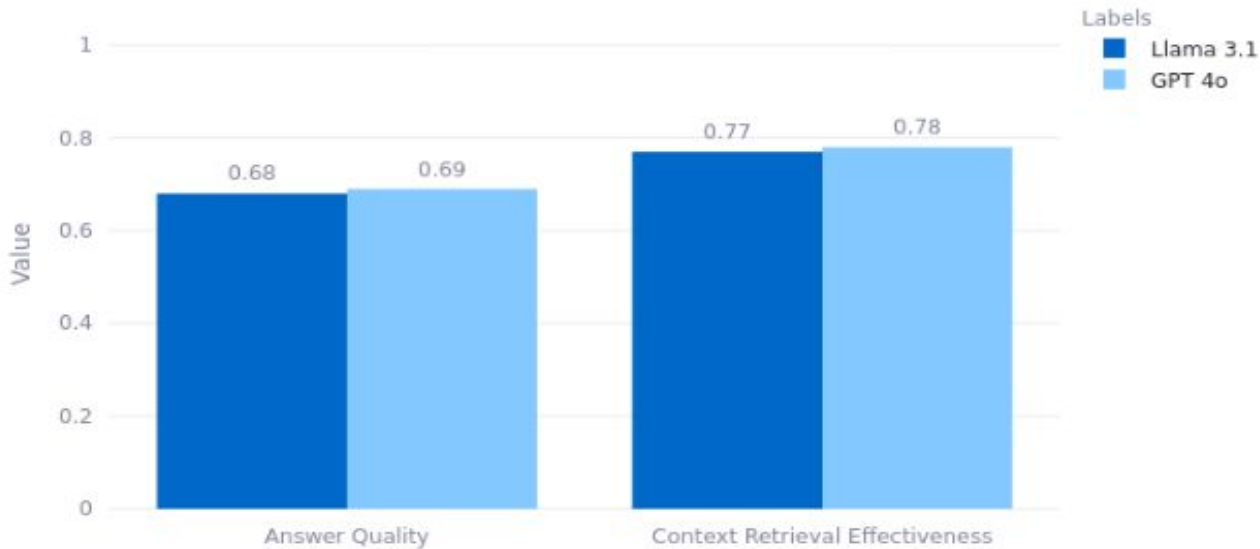
```
bash Copy Code
PUT _index_template/daily_logs
{
  "index_patterns": [
    "logs-*"
  ],
  "template": {
    "settings": {
      "number_of_shards": 2,
      "number_of_replicas": 1
    },
    "mappings": {
      "properties": {
        "timestamp": {
          "type": "date",
          "format": "yyyy-MM-dd HH:mm:ss|yyyy-MM-dd|epoch_millis"
        },
        "value": {
          "type": "double"
        }
      }
    },
    "priority": 1,
    "composed_of": [ "default_settings" ]
  }
}
```

The response continues: "This will create an index template named `daily_logs` that applies to index patterns matching the pattern `logs-*`. The template defines the settings and mappings for the index, including the number of shards and replicas, and the data type for the `timestamp` and `value` fields. You can then use the `GET _index_template` API to list all index templates, or the `GET _index_template/daily_logs` API to retrieve the configuration for the specific index template."

At the bottom, there is a link: `1 https://opensearch.docs.cern.ch/cluster_admin/index_templates/index.html`

Evaluation and Performance

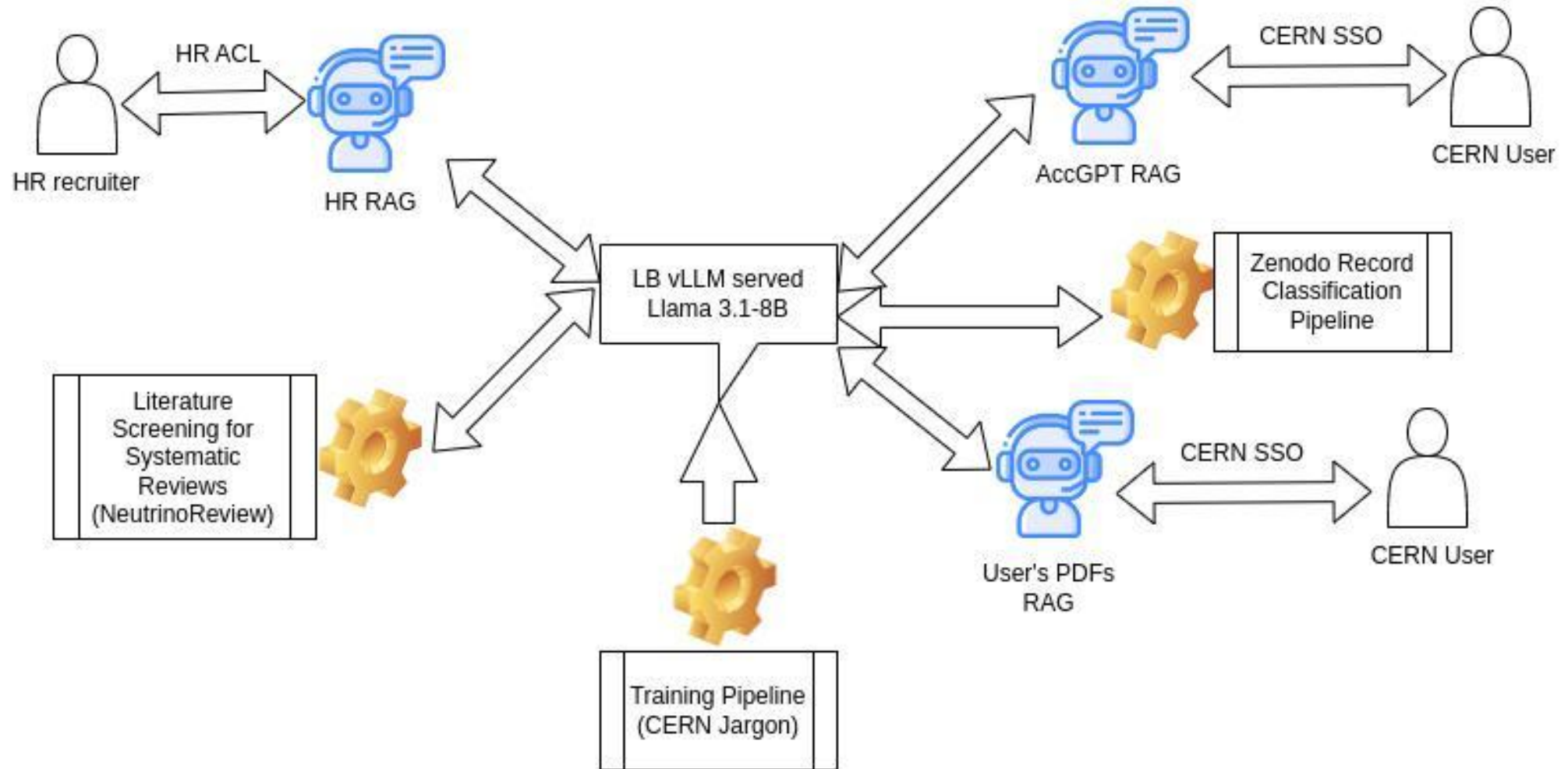
- **Metrics:**
 - Answer Quality (relevancy, correctness, faithfulness).
 - Context Retrieval Effectiveness (recall, precision, URL coverage, semantic similarity).
- **Evaluation Framework:** Ragas open-source framework.
- **Comparison:** AccGPT with Llama 3.1 vs. OpenAI GPT 4o (as a judge and a baseline).
- **Retrieved URL Coverage:** 92% accuracy in retrieving the correct URLs.
- **Context Retrieval Effectiveness:** 77-78%, demonstrating the accuracy of the RAG system.



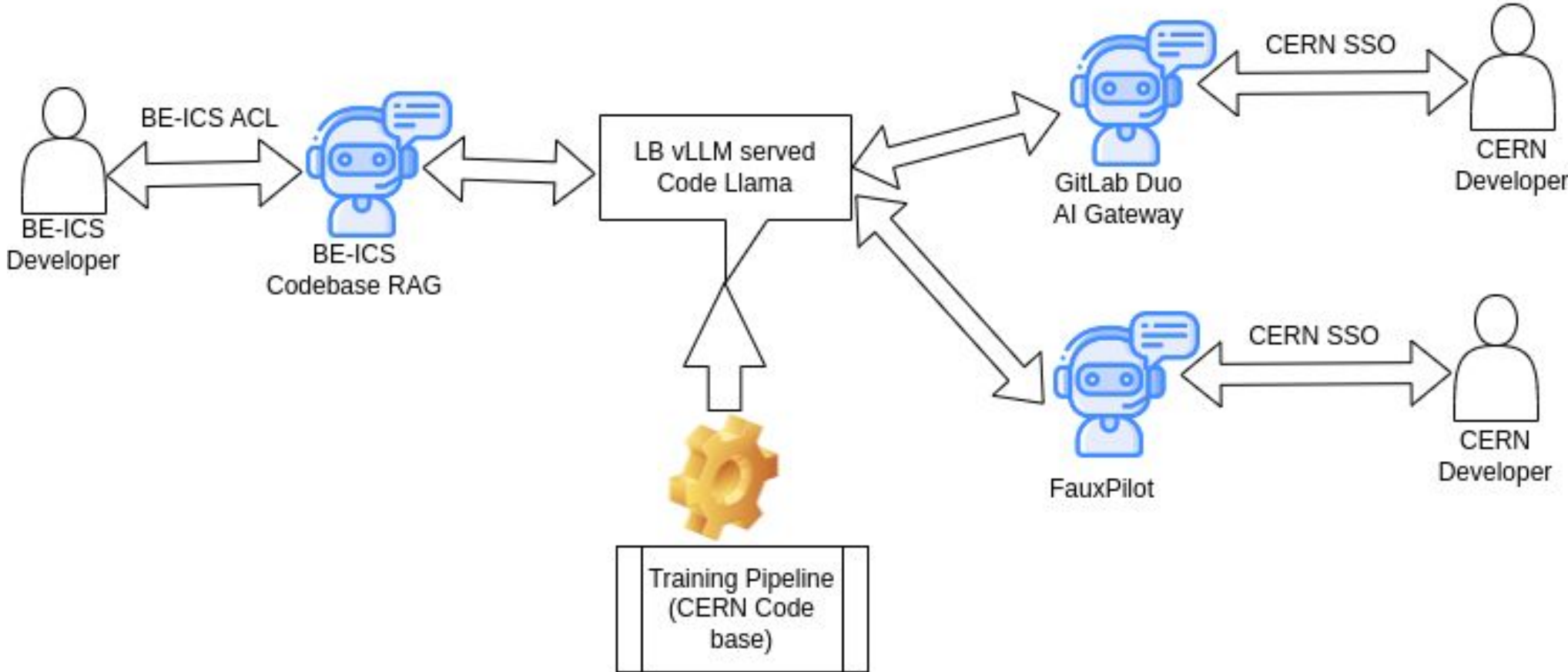
Lessons Learned

- **Internal Web Content Quality:** The importance of well-structured and accurate internal web content for effective RAG performance.
 - Knowledge Base growth can trigger wrong answers: e.g. things documented in several places (...but reference URL provided).
- **Open Source Community:** Embracing open-source solutions (like OpenWebUI) can save time and effort
 - Facilitates enormously AI Chatbots building
 - Can be used as single access point to several LLMs
 - Can be used to build RAG pipelines. OpenSearch RAG-Tool as well
- **Access Control Lists (ACLs):** The need for mechanisms to handle varying ACLs for different content within CERN's internal network
- **Right-sizing LLMs:** Selecting appropriate LLMs based on specific needs and optimizing their utilization to manage costs

Information Retrieval Use cases



Code Assistant Use cases



Conclusion

- AccGPT demonstrates the potential of **RAG-based LLM chatbots** for **efficient internal knowledge retrieval** in HEP sites and other organizations
- Documentation practices can be greatly improved, leading to increased efficiency and accuracy: It **identifies which internal documentation needs to be corrected**
- The success of LLM applications depends on:
 - A solid strategy
 - **Well-organized data**
 - Careful consideration of storage, compute options, and techniques (**moving target**)
- A Generative AI service is really needed at HEP sites:
 - To provide **access to LLMs**, and **tooling for AI application building**
 - To extend **the Service for Machine Learning lifecycle**
 - To also be usable for LLMs
 - To be based on OpenSource solutions (eg. <https://mlflow.org/>)
 - To provide a central **ML model repository**
 - To study the benefits and options for:
 - **Fine tuning LLMs on static knowledge**
 - **Per-Purpose AI Chatbots** rather than trying to compete with ChatGPT