

AccGPT - LLMs for CERN!

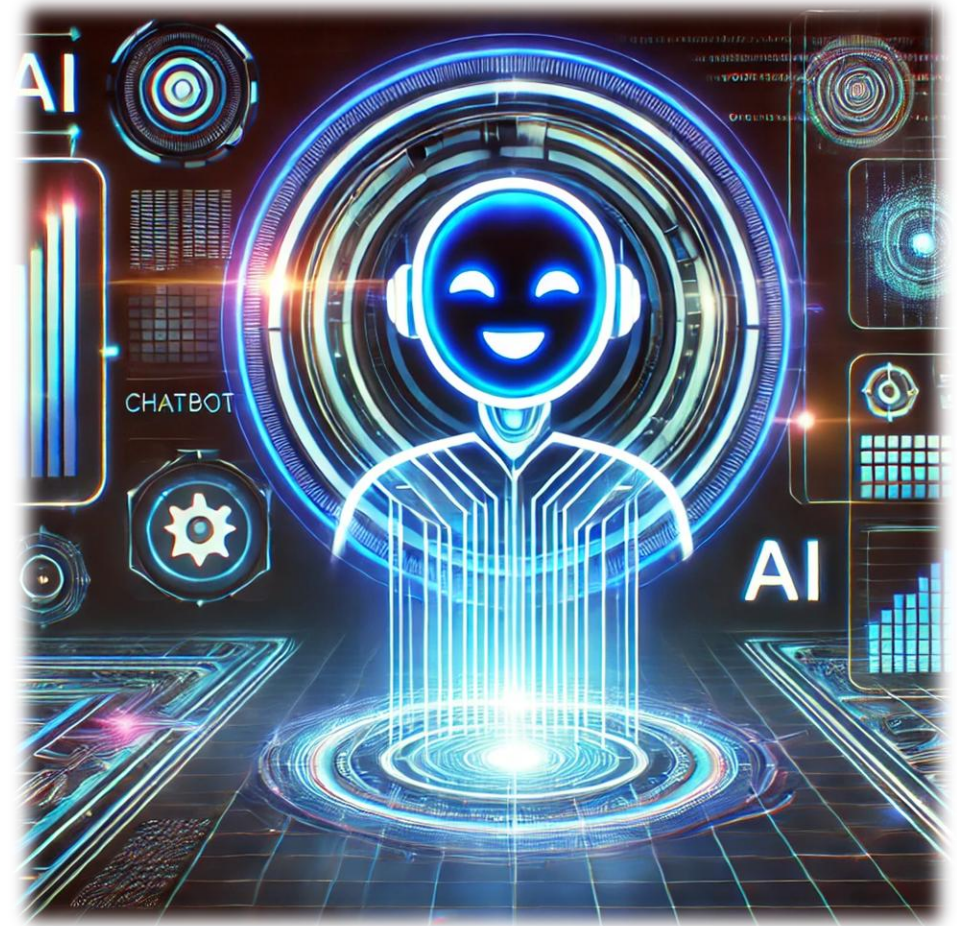
Vision and Use Case

Florian Rehm, Luke Van Leijenhorst, Verena Kain, Juan Manuel Guijarro, Sofia Vallecorsa

01.07.2025 - IML Machine Learning Working Group: Chatbots at CERN

Content

- What is AccGPT
- The AccGPT-Search model
- Current and future researches

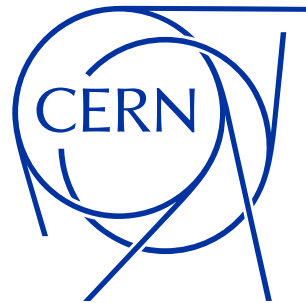


Introduction

AccGPT (Accelerating Science GPT).

- Our vision: Accelerating CERN research with AI.
- **AccGPT Pilot Service:** The entire service.
 - Providing LLMs to CERN researchers.
 - LLM access via UI (similar to ChatGPT) and API.
 - Assisting with LLM use case implementations.
- **AccGPT-Search Model:** Advanced expert-model.
 - An example use case: A chatbot for CERN internal knowledge search.

AccGPT Pilot LLM Service



AccGPT Pilot LLM Service

Goal:

- One central AI/LLM unit to provide access to LLMs (internal and external ones) and assist developers with expertise and usage examples.
 - Hosting and developing expert models.
 - Depending on resource dedication also consultancy.

Current Status:

- Still in pilot phase.
- CERN has not yet decided on further AI/LLM investment.
 - Though, support across all CERN hierarchy levels (next slide).
- Very limited resources (manpower and hardware).
 - PoC models are possible. Perfect production models difficult.

CERN Wants LLMs!

ATS-Sector-wide Survey:

- Participants well distributed across the entire ATS sector.
- 24 groups participated.

Central LLM Service at CERN

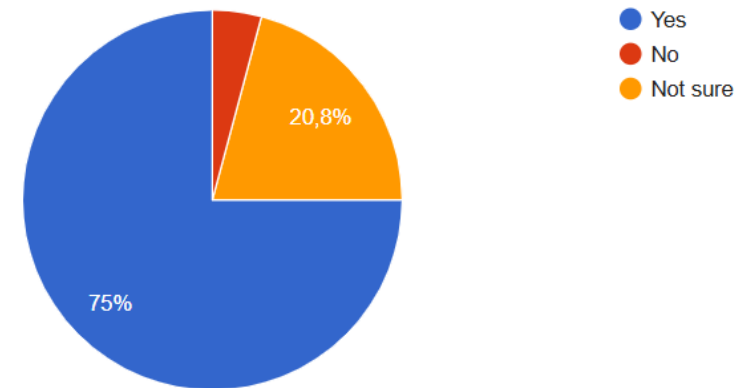
Do you think CERN should provide LLMs?

[Diagramm kopieren](#)

These could include advanced AI assistants interacting with CERN's documentation or assisting experts with tasks in the control room and in maintenance activities.

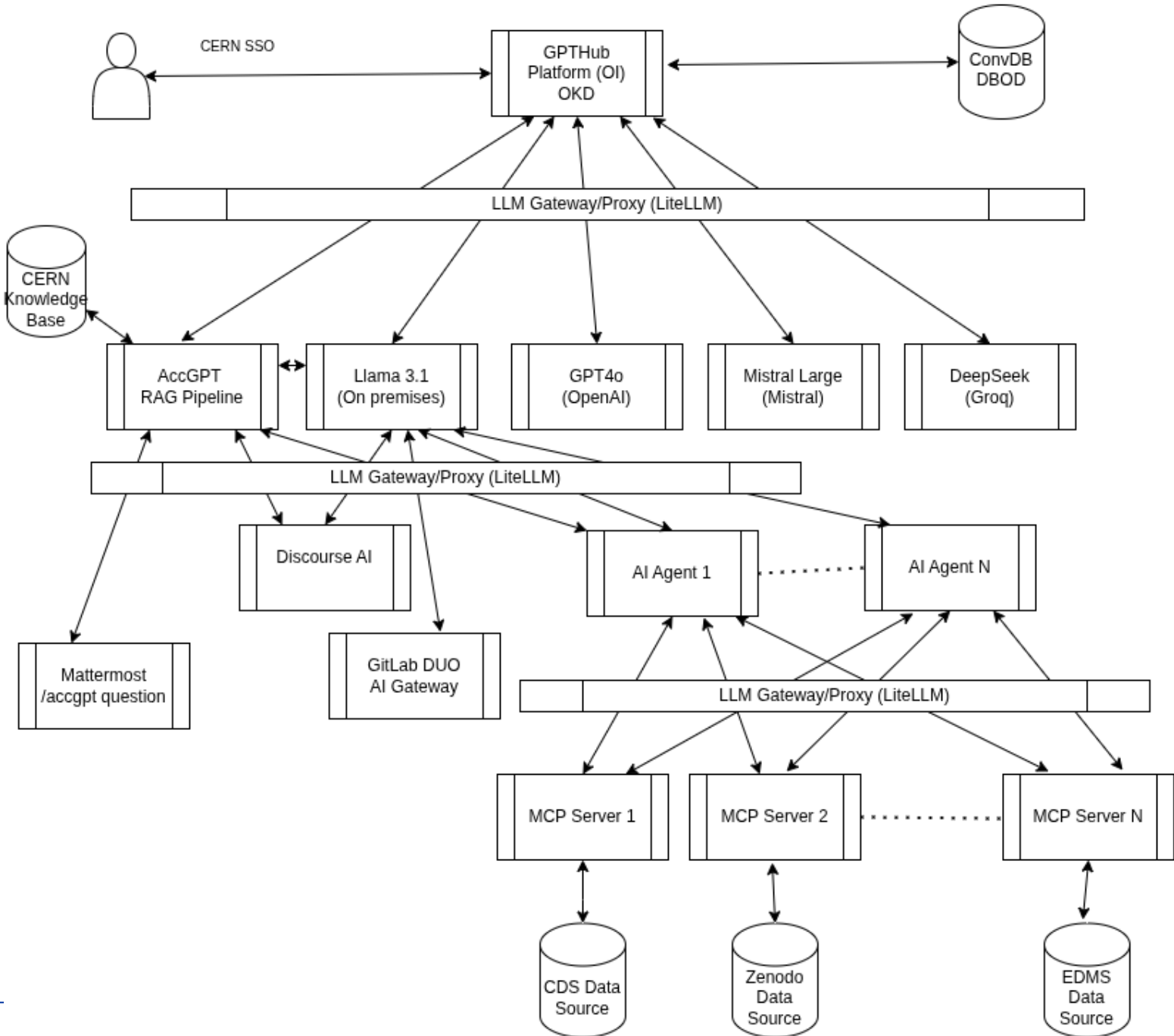
But also support general tasks such as coding or administrative work (e.g. writing minutes).

24 Antworten



→ **ATS users** expressed a clear interest in having a **central LLM service** for **expert and non-expert** use cases!

How Could it Look Like?

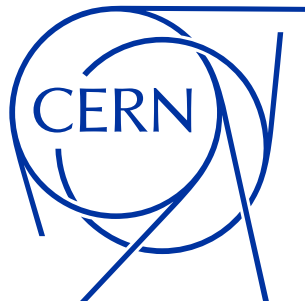


Draft LLM Gateway Platform

For brainstorming with and within the IT department.

By Juan Manuel Guijarro

Why AccGPT?



Why AccGPT?

A First PoC Model: Enhancing CERN knowledge retrieval.

- Challenge: **CERN has many and HUGE data bases:**
 - Challenging to find information without knowing its location.
 - >> 50 knowledge (web) domains for documentation.
 - Confluence (Wiki), CDS, CERN home, CERNbox, EDMS and more domains ...
 - Individual documentations.
 - Support tickets.



By GPT4

→ **Objective:** Leverage AccGPT to improve knowledge finding, user support, streamline development processes, and enhance onboarding experiences.

What Else?

Numerous software frameworks (such as FESA, UCAP, GeOFF, NXCALS, ...).

- With steep learning curves to use them.
→ AI assistance for learning and utilizing them.

Coding assistance:

- Many documentations contain code.
 - „Generate me a code template of ...“

Future AccGPT applications:

- Enhancements to machine and shift logbooks. ←
- Code Agents.

Future ++:

- AI assistant in the control room: conversational UI / controls.

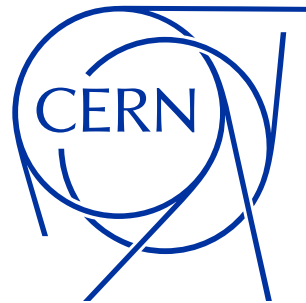
IPAC'24 - 15th International Particle Accelerator Conference

DESY. BERKELEY LAB Bringing Science Solutions to the World SLAC NATIONAL ACCELERATOR LABORATORY CERN Brookhaven National Laboratory Fermilab HELMHOLTZ HZB Helmholtz Zentrum Berlin

Towards Unlocking Insights from Logbooks Using AI

A. Sulc (HZB), G. Hartmann (HZN), J. Maldonado (BNL), V. Kain (CERN), F. Rehm (CERN), A. Eichler (DESY), J. Kaiser (DESY), T. Wilksen (DESY), F. Mayet (DESY), R. Kammering (DESY), H. Tuennemann (DESY), J. St. John (Fermi), H. Hoschouer (Fermi), K. J. Hazelwood (Fermi), T. Hellert (LBNL), D. Ratner (SLAC), W.-L. Hu (SLAC), A. Bien (SLAC)

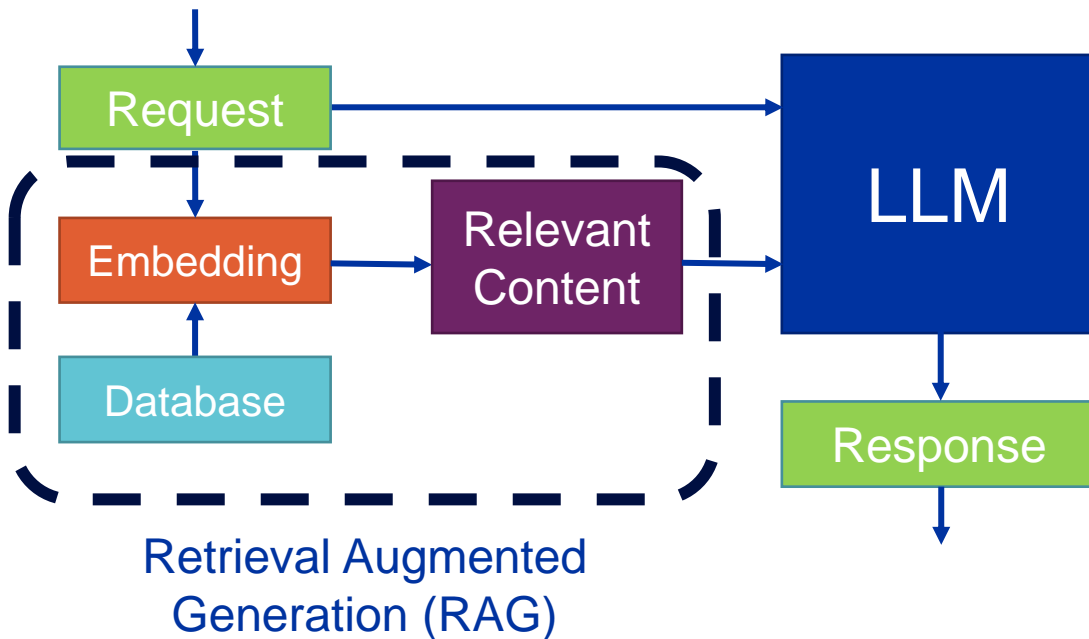
Inside of AccGPT-Search



Inside of AccGPT-Search – Conventional Setup

AccGPT-Search: A knowledge retrieval chatbot.

The (core) AccGPT pipeline*:



- **Based on two core models:**
 1. **Embedding model:**
 - Retrieves „relevant content“ from database.
 2. **Large Language Model (LLM):**
 - Formulates responses using the „relevant content“.
 - Focus on open-source models.
 - Possibility to use commercial models.
- **Accompanied by a self-created knowledge data base.**

* In reality the AccGPT pipeline has progressed and is more complex.

1. The Embedding Model

Function:

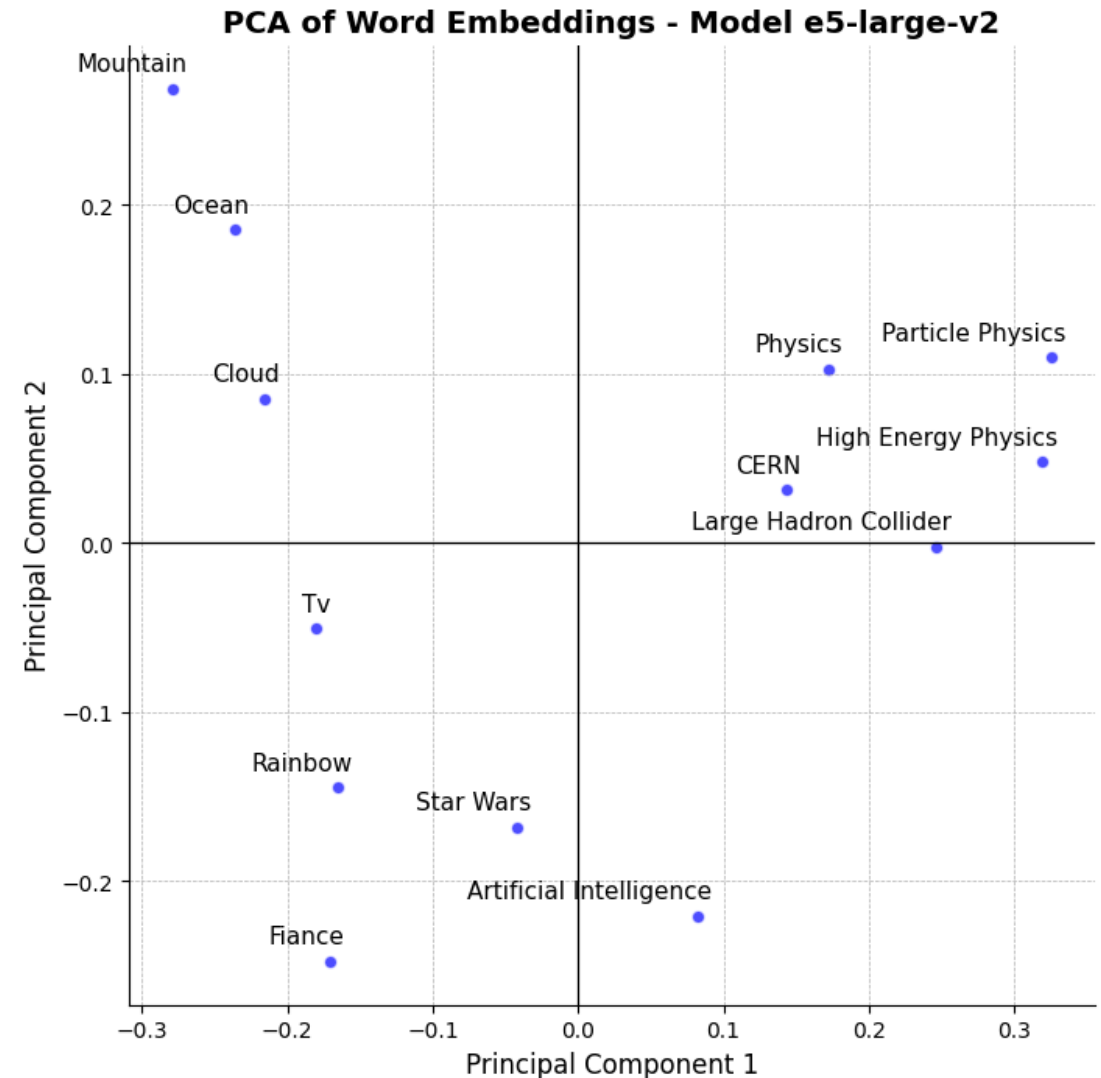
- Maps textual data into a latent space (1024 dimensional).
- Embed single words, sentences, or entire paragraphs.

Representation:

- Related phrases are close in the latent space.

In AccGPT-Search:

- Retrieve the most relevant content from our database based on user questions.
- This is done by taking the closest data point(s) to the question in the latent space.



2. The Large Language Model (LLM)

Definition:

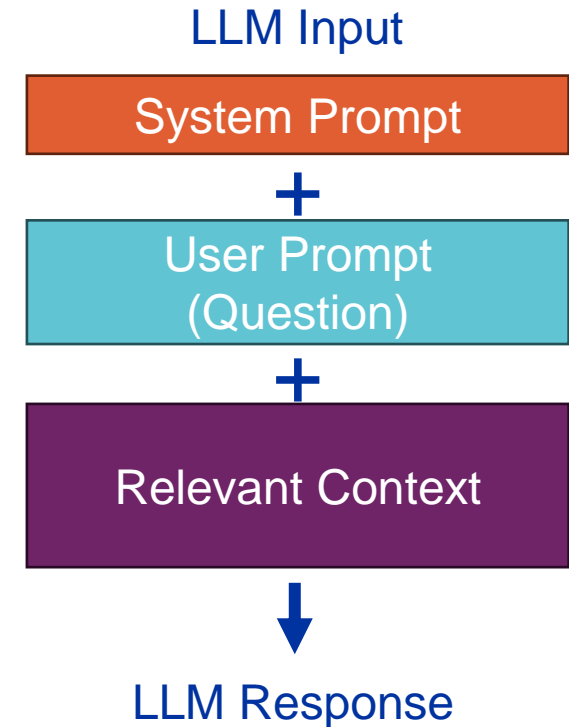
- Huge deep learning models trained on vast amounts of text to understand and generate human-like language.

Capabilities:

- Deep text understanding and context-awareness.

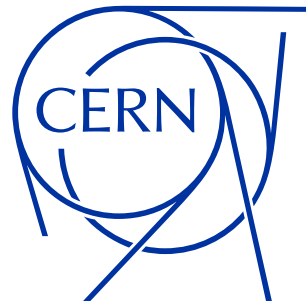
Our implementation:

- Using pre-trained open-source LLMs: **LLaMA**.
 - We (currently) don't fine-tune LLMs.
- Combined with external models, currently by Groq*.



*Groq designs ultra-fast AI inference hardware and hosts open-source LLMs via API, while Grok is a chatbot by xAI.

Some AccGPT Examples



The AccGPT User Interface

The screenshot displays the AccGPT web interface. At the top left, there is a menu icon followed by 'AccGPT' and a plus sign. Below this is a search bar labeled 'Search a model'. A dropdown menu is open, listing several models: 'gpt-4o', 'o1-mini', 'o1-preview', 'Llama-70B (Groq)', 'deepseek-r1 (Groq)', 'mixtral-8x7b (Groq)', and 'Llama-8B (local)'. The 'gpt-4o' model is highlighted with a blue rounded rectangle. To the right of the dropdown, the AccGPT logo is visible, with a blue arrow pointing to it. Below the logo, the text 'AccGPT 0.0.49 RAG on CERN web domains' is displayed. At the bottom of the interface, there is a chat input area with the placeholder text 'How can I help you today?'. Below the input area, there are icons for '+', 'Image', and 'Code Interpreter'. A 'Suggested' section is visible below the input area, containing the question 'What are the supported Linux distributions at CERN?' and the answer 'Linux at CERN'.

AccGPT: Example 1

The screenshot shows a chat window with a header containing 'AccGPT' and 'gpt-4o'. A message input field at the top right contains the text 'What is AccGPT?'. The chat history shows two messages:

- AccGPT:** AccGPT is a chatbot pilot project at CERN that utilizes Large Language Models (LLMs) to interact with CERN's internal knowledge base, providing quick and straightforward answers to queries, similar to ChatGPT. It includes two source links: <https://confluence.cern.ch/display/BECSSD...> and <https://beams....>
- gpt-4o:** AccGPT is not a widely recognized term or concept as of my last update in October 2023. It's possible that it could be a specialized tool, a software product, a project, or an initiative developed after that date, or it might be a term used in a specific niche community. If you have additional context or details about what AccGPT refers to, I might be able to provide more targeted information. Otherwise, checking the latest sources or the context in which you encountered the term could provide more up-to-date insights.

AccGPT: Example 2

Which web browsers are supported at CERN?

OI accgpt

Microsoft Edge is the recommended web browser at CERN for Windows. Additionally, Firefox and Google Chrome are also supported on all platforms.

1 https://cern.service-now.com/service-portal?id=kb_article&n=KB0006225



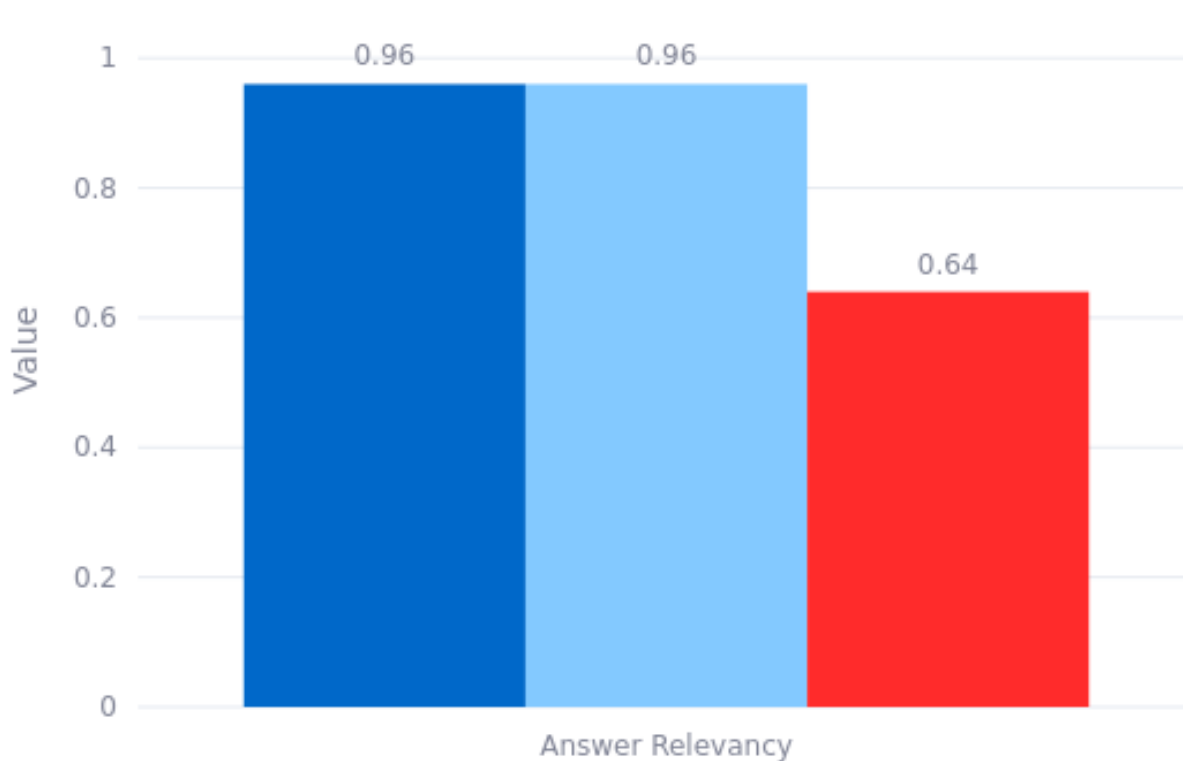
Feedback

References

Disclaimer:

We do not guarantee the accuracy or completeness of the provided responses. We strongly encourage users to verify the information.

Accuracy: RAG vs non-RAG



Labels

- AccGPT (GPT 4o)
- AccGPT (Llama 3.1 8B)
- GPT 4o (no RAG)

(RAG: Retrieval Augmented Generation)

With RAG (AccGPT):

- **AccGPT performs extremely well**, independently of the LLM.
- For Q&A set, RAG retrieves the correct knowledge in **>90%** of cases (not shown).

Without RAG:

- Accuracy drops, as **default LLM is not aware of CERN knowledge**.
- Not zero, since many questions are not exclusive to CERN.

Answer Relevancy:

Evaluates how accurately the LLM response matches experts' questions on CERN related topics.

(Curated dataset from experts)

Data Quality Importance

Not only AccGPT determines the accuracy:

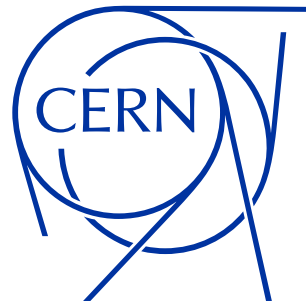
- **Data quality is KEY !!!**
(garbage in ... garbage out)
- Important aspects:
 - Keep documentations updated.
 - Keep documentations detailed.
 - Duplicates make the model struggling.
 - Text based information is best.
 - Ideally: FAQs and examples.
- For coding:
 - Templates or examples make it easier for the model.



By GPT 4

→ Keep this in mind if you plan to make your documentation accessible to chatbots.

New Feature



AccGPT in Mattermost !!!

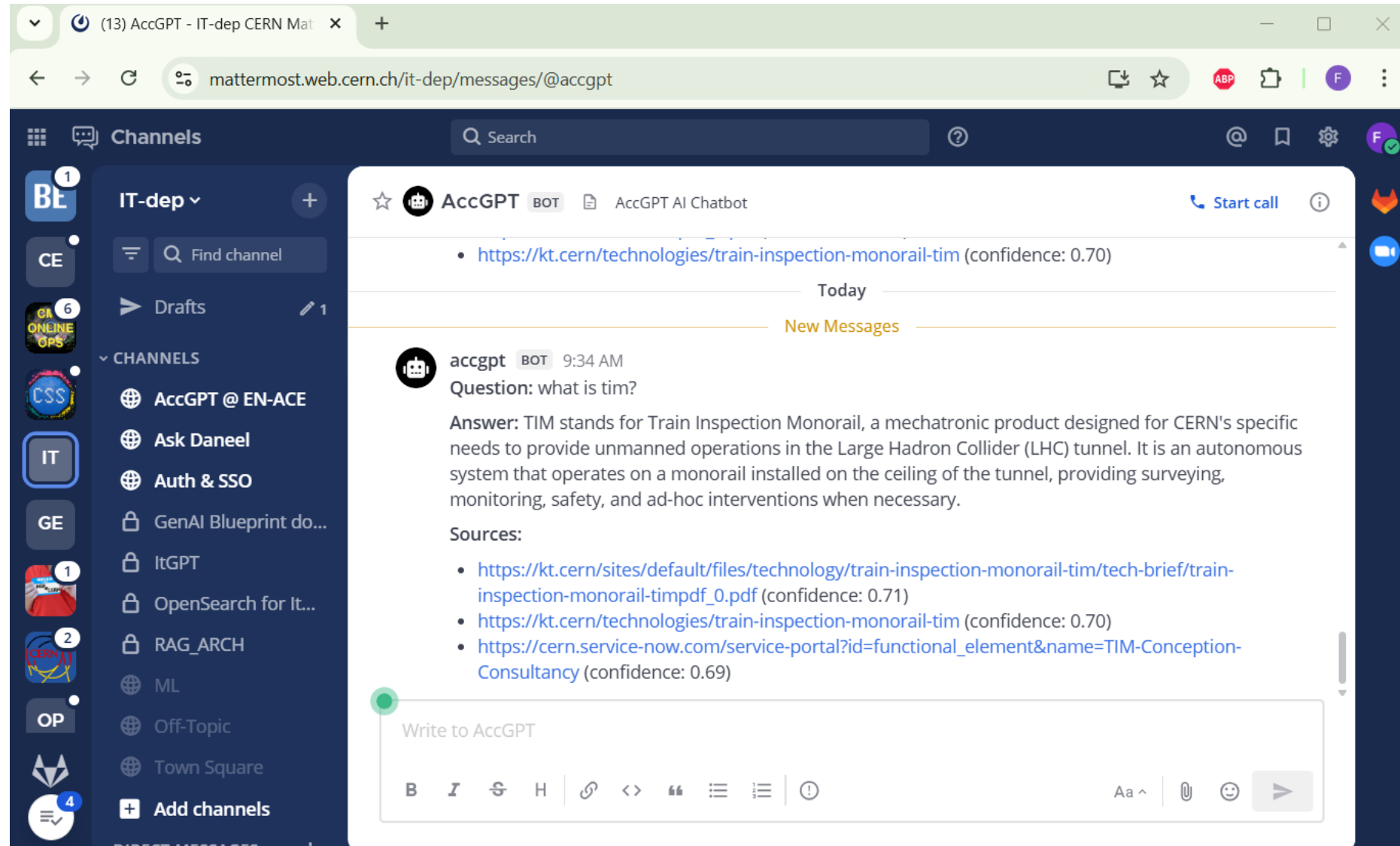
Usage:

/accgpt [message]

Where?

- Dedicated chat.
- Within every chat.

- *Sometimes buggy...*
- *Still older AccGPT version.*



The screenshot shows a Mattermost chat window in a browser. The address bar indicates the URL is `mattermost.web.cern.ch/it-dep/messages/@accgpt`. The chat interface shows a message from the **AccGPT BOT** (AccGPT AI Chatbot) at 9:34 AM. The message content is:

Question: what is tim?

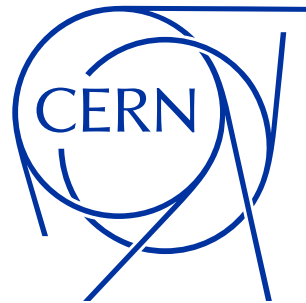
Answer: TIM stands for Train Inspection Monorail, a mechatronic product designed for CERN's specific needs to provide unmanned operations in the Large Hadron Collider (LHC) tunnel. It is an autonomous system that operates on a monorail installed on the ceiling of the tunnel, providing surveying, monitoring, safety, and ad-hoc interventions when necessary.

Sources:

- <https://kt.cern/technologies/train-inspection-monorail-tim> (confidence: 0.70)
- https://kt.cern/sites/default/files/technology/train-inspection-monorail-tim/tech-brief/train-inspection-monorail-timpdf_0.pdf (confidence: 0.71)
- <https://kt.cern/technologies/train-inspection-monorail-tim> (confidence: 0.70)
- https://cern.service-now.com/service-portal?id=functional_element&name=TIM-Conception-Consultancy (confidence: 0.69)

The chat interface also shows a list of channels on the left, including IT-dep, CE, CA ONLINE OPS, CSS, IT, GE, and OP. The chat input field at the bottom contains the text "Write to AccGPT" and has a rich text editor with various formatting options.

Next Steps



AccGPT: Next Steps - Agents

```
In [5]: agent_executor.invoke({"input": "What is the phone number of Florian Rehm and where is his office"})
```

```
> Entering new AgentExecutor chain...
```

```
Invoking: `CERN_phonebook_search` with `{'query': 'Florian Rehm'}`
```

```
Preferred name: Florian REHM  
Organic Unit: BE-CSS-DSB  
Email: florian.matthias.rehm@cern.ch  
Phone number: 66134; fullNumber: "+41227666134"  
Office location: 774/2-030  
Mail box: Z10400
```

```
> Finished chain.
```

```
Out[5]: {'input': 'What is the phone number of Florian Rehm and where is his office',  
         'output': "Florian Rehm's phone number is +41227666134 and his office is located at 774/2-030."}
```

- Agents can do sub-tasks by calling functions.
- For example:
 - Querying the CERN phonebook.
 - However: The phonebook feature will not come (soon), due to data privacy reasons.

AccGPT-Agent

- **Making AccGPT agentic:**
 - **Normal RAG:** simple single-step retrieval workflow.
 - **Agentic RAG:** multi-step adaptive workflow.
- **Strength:**
 - Answers are more detailed and complete.
- **Weakness:**
 - Sequential process, slower than AccGPT, more LLM calls required.
- Currently the only tool is RAG search. More tools to come.

Example Question:

*What were **the major activities and upgrades** carried out **during each of CERN's Long Shutdowns**?*

Step	Agentic retrieval steps
1	CERN Long Shutdowns
2	Timeline Long Shutdown
3	LS1 activities and upgrades
4	LS2 activities and upgrades
5	LS3 activities and upgrades

AccGPT-GraphRAG

Utilization of a Knowledge Graph instead of simple RAG.

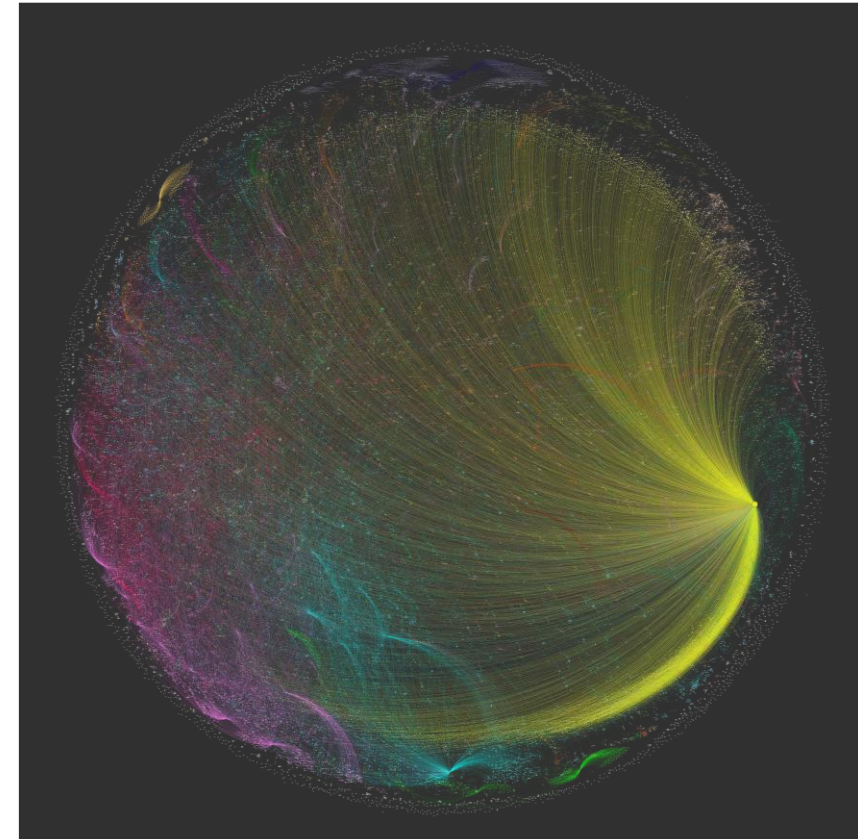
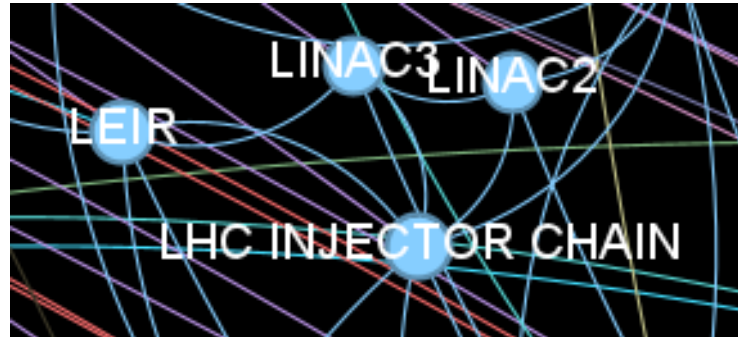
- **Knowledge Graph:** A network of connected entities and relationships that captures structured information.
 - Graph generated by LLMs.

Strengths:

- Enables precise retrieval across complex, interlinked data structures.
- Generates high-level, coherent overviews across large knowledge domains.

Weaknesses:

- Graph creation is expensive and time-consuming (~1 million LLM calls).



Coding Agents

- Coding Agents differ from standard LLM Agents by using programmatic reasoning and executing complete code workflows, rather than relying on step-by-step API calls.

Generate a table of the settings of the parameter group TUNE_TRIM in LHC BP for the beam process name 'PHYSICS-6.8TeV-1.2m-2024_V1@135_[END]'

```
— Executing parsed code: —
import pandas as pd

# Step 1: Verify the beam process exists
beam_process_name = 'PHYSICS-6.8TeV-1.2m-2024_V1@135_[END]'
beam_processes = find_beamprocesses(accelerator='LHC',
name=beam_process_name)
print(beam_processes)

# Step 2: Get parameters in the group TUNE_TRIM
parameters = find_parameters(accelerator='LHC', group=['TUNE_TRIM'])
print(parameters)

# Step 3: Retrieve settings for the beam process and parameters
settings = find_settings(beamprocess_name=beam_process_name,
parameters=parameters)
print(settings)

# Step 4: Convert the settings to a table and provide the final answer
settings_df = pd.DataFrame(settings)
final_answer(settings_df)
```



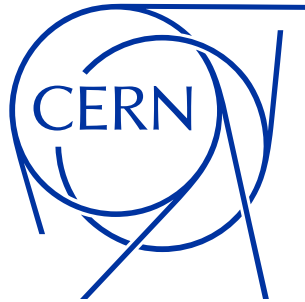
	Parameter	Value
0	RPMBB.UA27.RQTD.A23B1/IREF	-76.691700
1	RQTF.A78B2/I	-3.392400
2	LHCBEAM2/QH_TRIM_FIDEL	0.000000
3	RQTD.A34B2/K1	-0.001426
4	RQTD.A81B1/K_SMOOTH	0.000308
...
131	RQTD.A56B2/K1	0.000333
132	RQTD.A34B2/I	134.740560
133	RQTF.A78B2/K_SMOOTH	-0.000036
134	RPMBB.UA63.RQTD.A56B1/IREF	16.221600
135	RQTF.A23B1/K1	-0.000090

Embedding Model Fine-Tuning

- **Fine-tuning a full LLM** → (currently) **not feasible (hardware constraints)**.
- **Fine-tuning an embedding model** → **feasible**:
 - Much smaller (hundreds of millions of parameters).
 - For AccGPT-search (RAG):
 - Accurate embedding retrieval is key for high performance.
 - Inspired by ChATLAS:
 - Building a Q&A dataset for fine-tuning.

Work in progress.

Summary



Summary

- **AccGPT:** A CERN LLM service pilot.
- **AccGPT-Search:** A PoC model for knowledge retrieval.
 - Good accuracy, growing capabilities.
 - Achieved a lot with limited resources.
 - Much potential, but lacking production-scale support.
- Many developments and researches in the pipeline.
- **AccGPT Resources:**
 - [Knowledge sources.](#)
 - [Landing page & registration.](#)
 - [User interface.](#)

