AccGPT

A Chatbot for CERN Internal Knowledge

Florian Rehm, Verena Kain, Juan Manuel Guijarro, Sofia Vallecorsa

20.09.2024

Content

- What is AccGPT.
- How does it work.
- Some examples.
- Who is behind.
- Future steps.

What is AccGPT?

AccGPT Definition

AccGPT (Accelerating GPT).

- Our vision: Accelerating CERN research with AI.

A chatbot for CERN-specific knowledge:

- Based on Large Language Models (LLMs).
- Enhanced with Retrieval Augmented Generation (RAG).
  - Integration of a CERN-specific knowledge database.
- More details later.

→ Goal: Simplify and accelerate the process of finding CERN information.

Why AccGPT?

Why AccGPT?

By GPT4

First step: Enhancing CERN knowledge search.

- **Challenge:** CERN has many and HUGE data bases:
  - >> 50 knowledge (web) domains for documentation.
    - Challenging to find information without knowing its location.
  - CERN Document Server (CDS): > 500k documents.
  - CERN home: > 10k webpages.
  - Confluence (Wiki)
  - CERNbox
  - And many more domains …

→ **Objective:** Leverage AccGPT to improve knowledge finding, user support, speed-up development processess and enhance onboarding experiences.

Why AccGPT? The List Goes on …

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 (internal) GitLab repositories.
- Aid in code development, understanding and debugging.

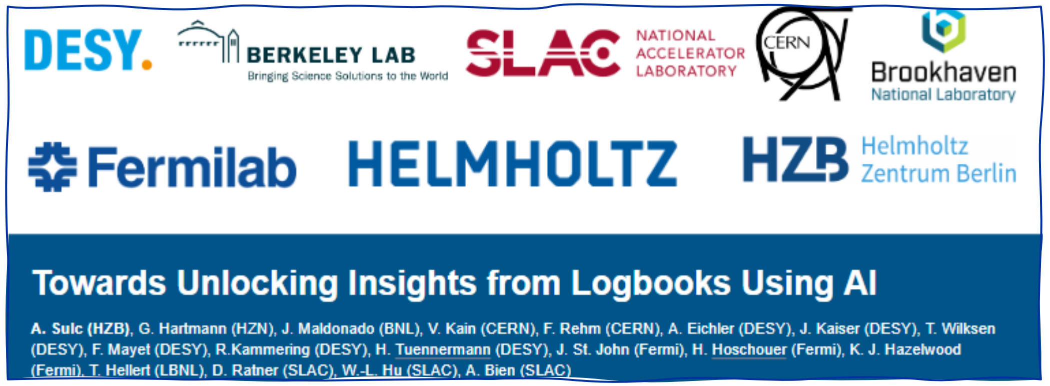
Future AccGPT applications:

- Enhancements to machine and shift logbooks.

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

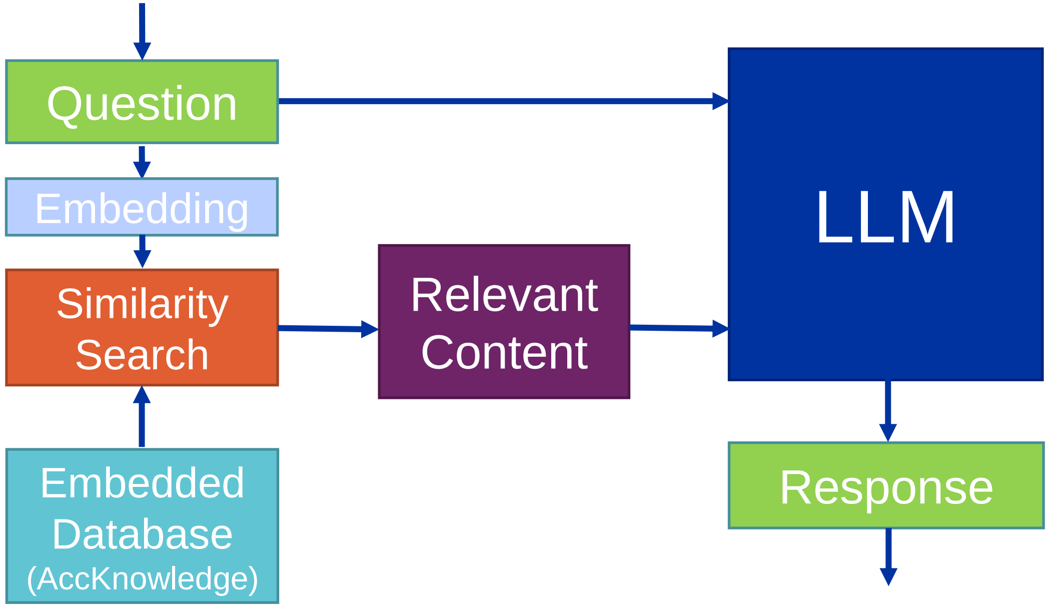
Towards Unlocking Insights from Logbooks Using AI

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. Tuennermann (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

Inside of AccGPT

The (core) AccGPT pipeline*:
Retrieval Augmented Generation (RAG).



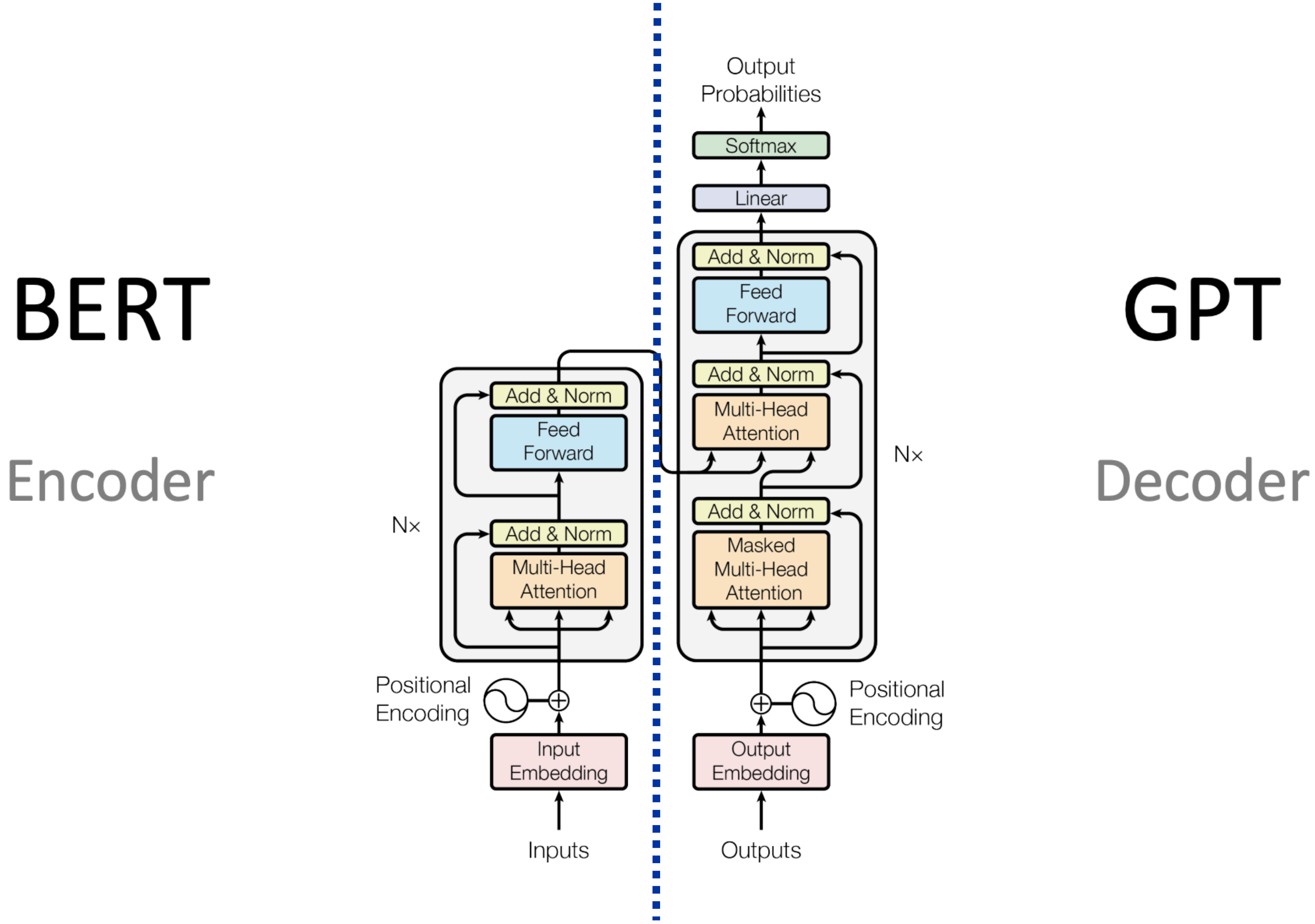
- **Based on two core models:**
  1. Embedding model:
     - Retrieves „relevant content“ from database.
  2. Large Language Model (LLM):
     - Formulates responses using the „relevant content“.
- **Accompanied by a self-created knowledge data base.**

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

Based on Transformer Architecture

The transformer architecture:

- Encoder part:
  - Used for embedding.
- Decoder part:
  - Used for generation.



DOI: 10.54851/v6i1y202401

1. The Embedding Model

Function:

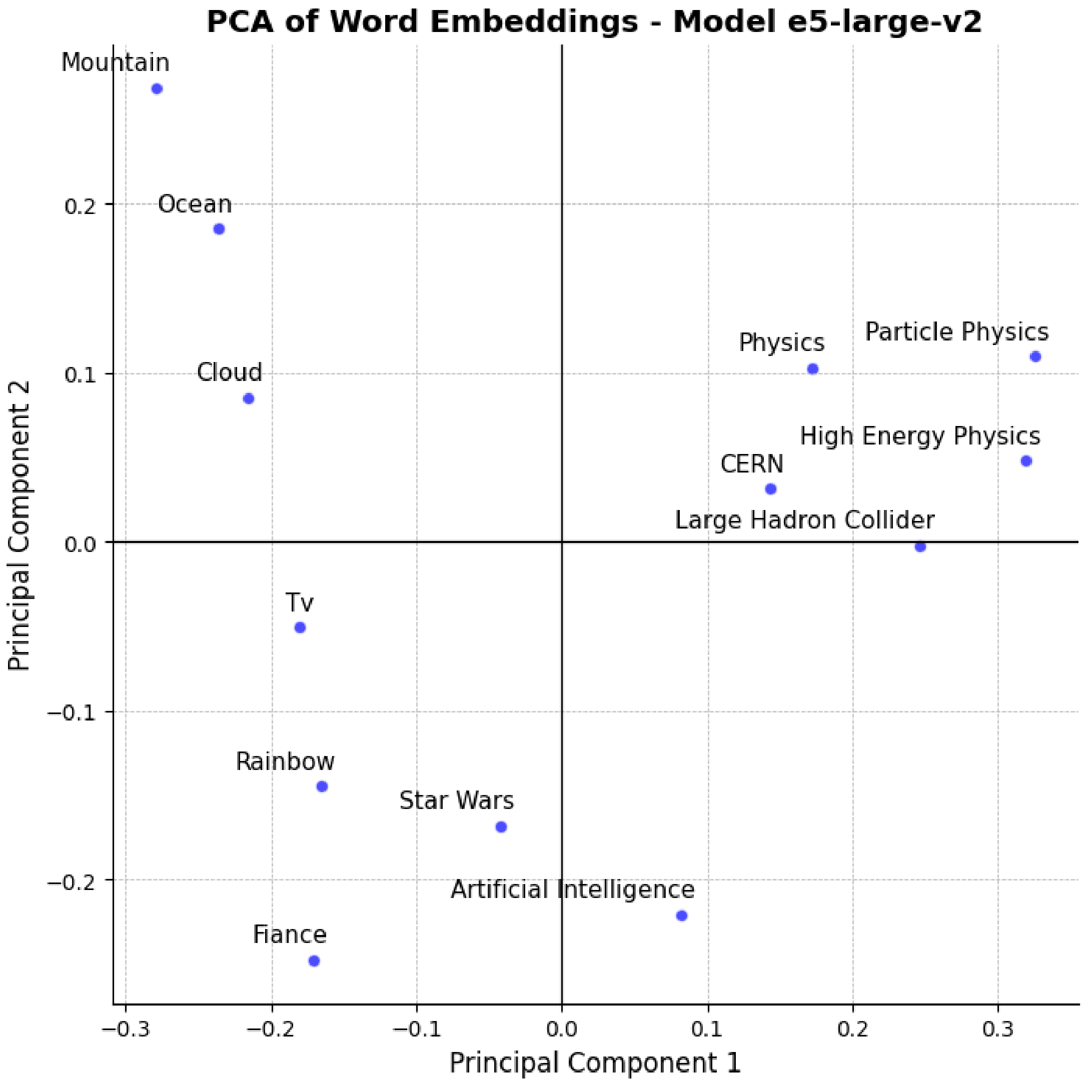
- Maps textual data into a latent space (1024 dimensional).
- Can embed single words, sentences, or entire paragraphs.

Representation:

- Related phrases are close in the latent space.

Our application:

- Retrieve the most relevant content from our database based on user questions.
- Taking the closest data point(s) to the question in the latent space.
- Utilizing open-source embedding models:
  - In use: BERT* e5-large-v2
  - Previous semantic-search models: Word2Vec, GloVe, FastText, …



* Bidirectional Encoder Representations from Transformers. EmbEddings from bidirEctional Encoder rEpresentations (https://arxiv.org/pdf/2212.03533.pdf)

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.

Application:

- Chatbots, translation, summarization, extraction, ...

LLM Input

System Prompt

+

User Prompt (Question)

+

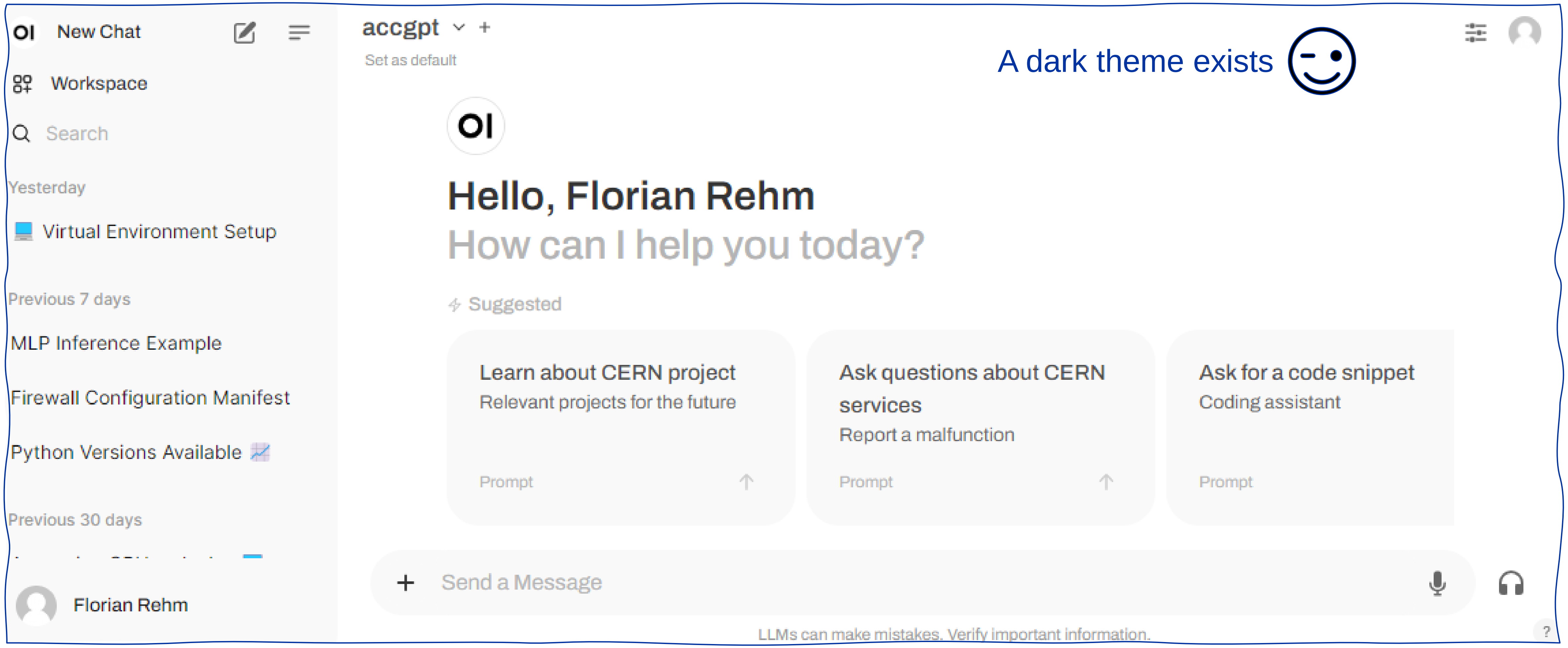
Relevant Context

Our implementation:

- Using pre-trained open-source LLMs: **LLaMA 3.1 8B.**

Some AccGPT Examples

The New User Interface!



New Chat

Workspace

Search

Yesterday

💻 Virtual Environment Setup

Previous 7 days

MLP Inference Example

Firewall Configuration Manifest

Python Versions Available 📈

Previous 30 days

Florian Rehm

accgpt

Set as default

A dark theme exists 😉

Hello, Florian Rehm

How can I help you today?

Suggested

Learn about CERN project
Relevant projects for the future
Prompt

Ask questions about CERN services
Report a malfunction
Prompt

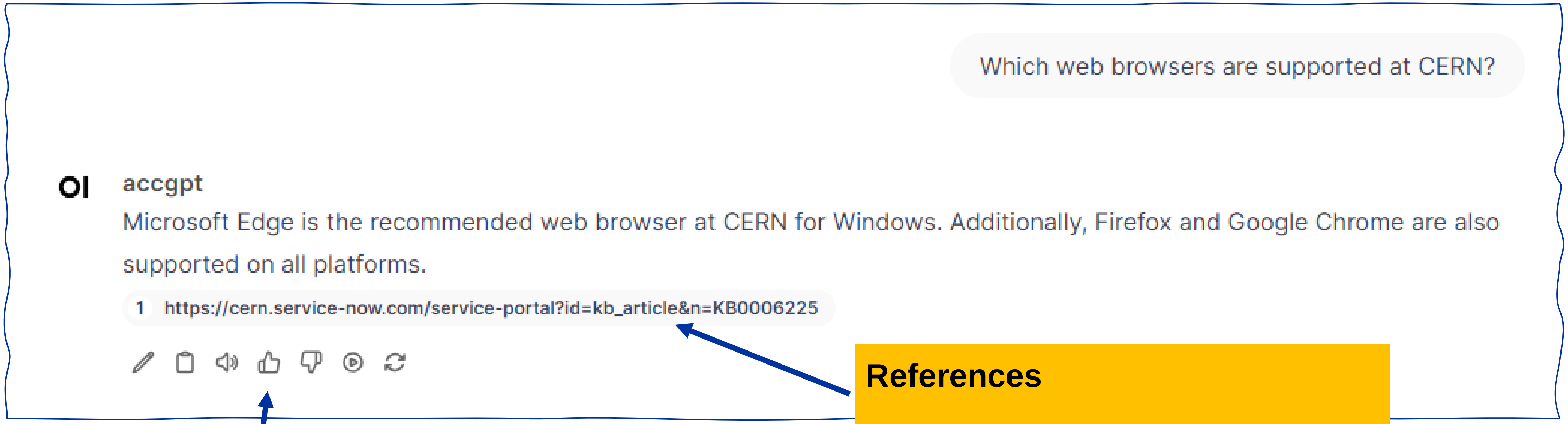
Ask for a code snippet
Coding assistant
Prompt

Send a Message

LLMs can make mistakes. Verify important information.

- **Open-source UI: Open WebUI**

AccGPT: Examples 1



Which web browsers are supported at CERN?

accgpt

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

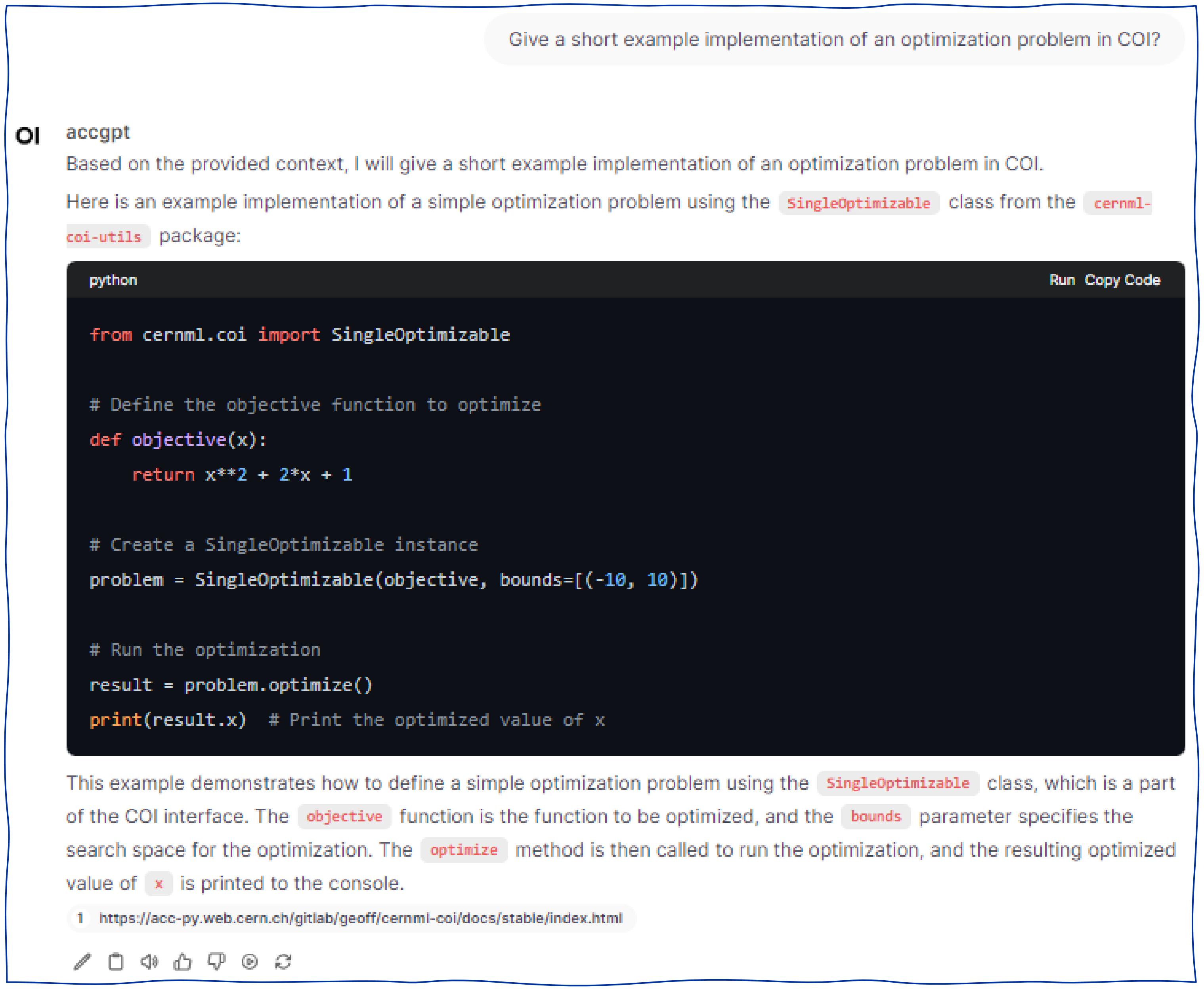
1 https://cern.service-now.com/service-portal?id=kb_article&n=KB0006225

Feedback

References

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

AccGPT: Examples 2



Give a short example implementation of an optimization problem in COI?

accgpt

Based on the provided context, I will give a short example implementation of an optimization problem in COI.

Here is an example implementation of a simple optimization problem using the `SingleOptimizable` class from the `cernml-coi-utils` package:

```python
from cernml.coi import SingleOptimizable

# Define the objective function to optimize
def objective(x):
    return x**2 + 2*x + 1

# Create a SingleOptimizable instance
problem = SingleOptimizable(objective, bounds=[(-10, 10)])

# Run the optimization
result = problem.optimize()
print(result.x)  # Print the optimized value of x
```

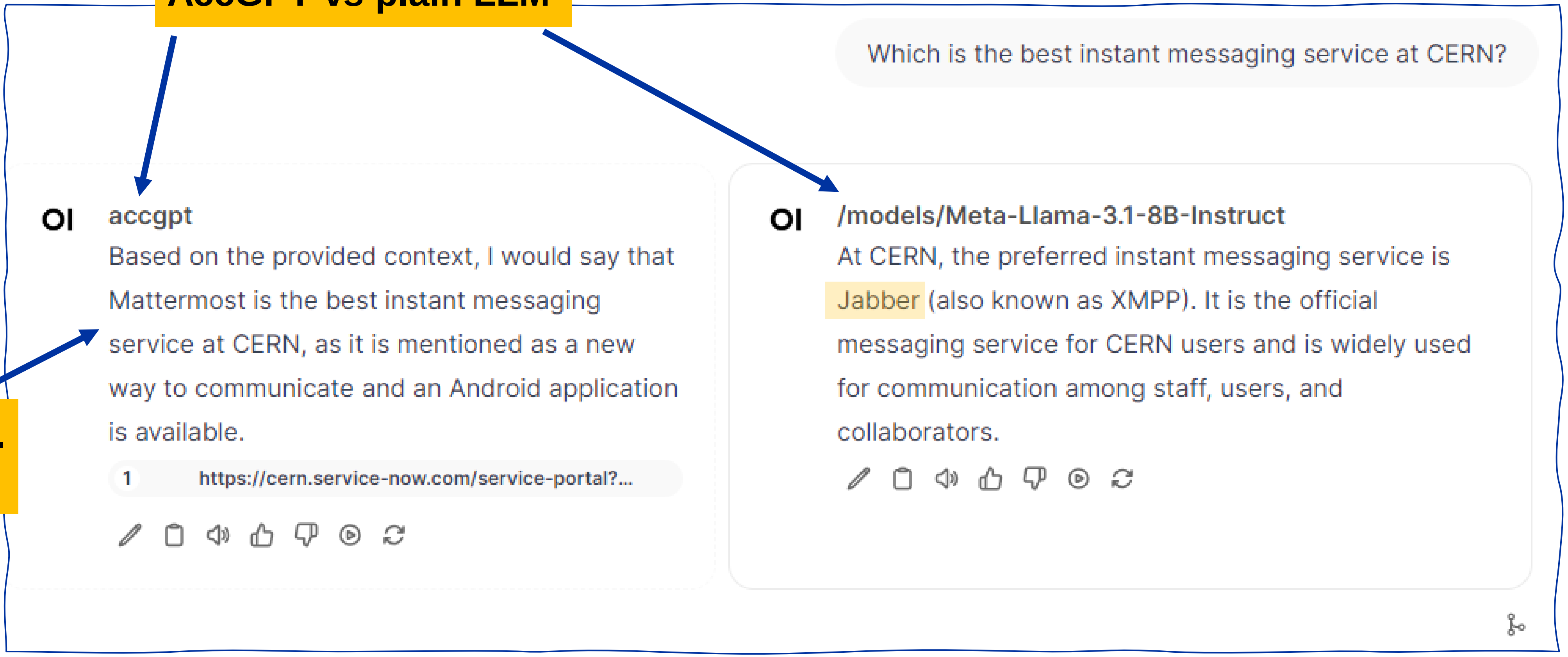
This example demonstrates how to define a simple optimization problem using the `SingleOptimizable` class, which is a part of the COI interface. The `objective` function is the function to be optimized, and the `bounds` parameter specifies the search space for the optimization. The `optimize` method is then called to run the optimization, and the resulting optimized value of `x` is printed to the console.

1 https://acc-py.web.cern.ch/gitlab/geoff/cernml-coi/docs/stable/index.html

For such tasks we aim to implement a dedicated coding LLM.

AccGPT: Examples 3

AccGPT vs plain LLM



Which is the best instant messaging service at CERN?

accgpt

Based on the provided context, I would say that Mattermost is the best instant messaging service at CERN, as it is mentioned as a new way to communicate and an Android application is available.

1 https://cern.service-now.com/service-portal?...

/models/Meta-Llama-3.1-8B-Instruct

At CERN, the preferred instant messaging service is Jabber (also known as XMPP). It is the official messaging service for CERN users and is widely used for communication among staff, users, and collaborators.

Correct answer.
Phrasing ok.

Data Quality Importance

Not only AccGPT determines the accuracy:

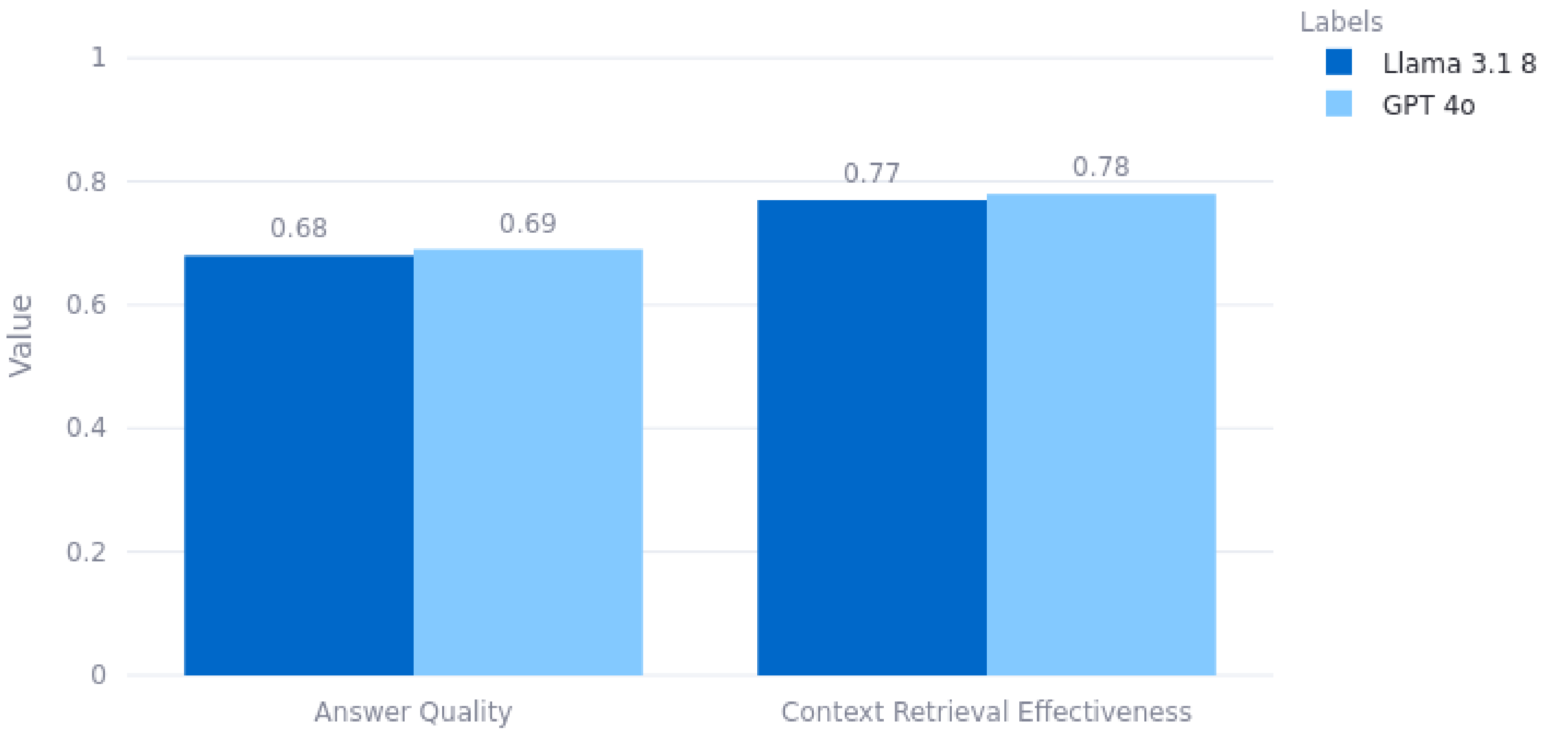
- **Data Quality is KEY !!!**
  (garbage in … garbage out)
- Important aspects:
  - Keep documentations updated.
  - Keep documentations detailed.
  - Duplicates make the model struggling.
  - Text based information is best.
  - Ideally: FAQs and examples.
- For coding:
  - Templates or examples make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

By GPT 4

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

AccGPT Evaluation

Comparing Llama vs GPT



→ AccGPT performs quite good.
→ Better LLMs do not improve accuracy.

Objective evaluation of generative models is tricky…

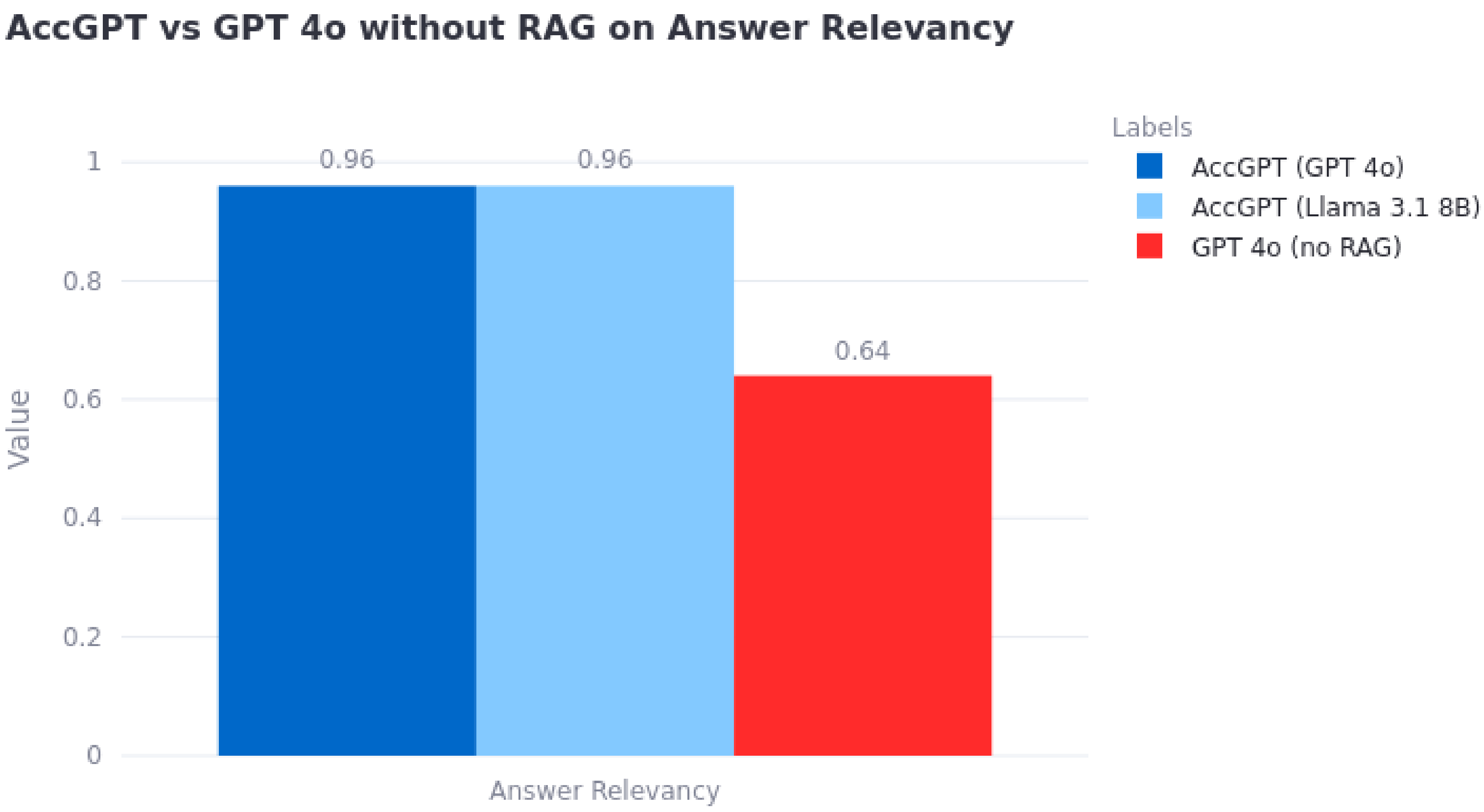
Two self-defined metrics:

- **The Answer Quality** measures how accurate the responses are compared to the True answers from experts.
- **The Context Retrieval Effectiveness** indicates the RAG system context retrieval accuracy.

Dataset: ~100 from experts provided Q&A pairs.

RAG vs non-RAG

- **Answer Relevancy:**
  - Evaluates how accurately the LLM response matches the user's question.
- **With RAG:**
  - AccGPT performs extremely good, independently from the LLM.
- **Without RAG:**
  - Worse accuracy, as LLM is not aware of CERN content.



AccGPT: A Collaborative Effort

Collaborative Effort Across Departments

Very limited resources for this project.
- ➢ Progress very slow

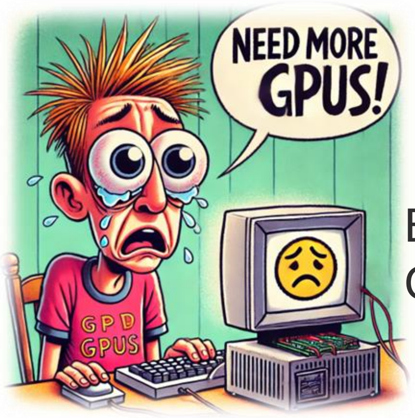
Key contributors:
- IT Department:
  - Infrastructure management and model hosting (0.5 FTEs).
- BE Department:
  - Development of the chatbot model (1 FTE).
    - Assistance of 2 summer students.

Additional support from volunteers:
- Contributions from some CERN members on a volunteer basis.

Goal:
- Provide AccGPT for entire CERN.
  - Minimize redundant work efforts across teams on similar projects.

Next Steps

AccGPT: Next Steps



By
GPT4

Improving AccGPT…

Scale up GPU resources:

- More GPUs allow usage of larger and better models.
  - Smaller models: “hallucinate” often, do not allow a chat-like conversation, are prompt sensitive.

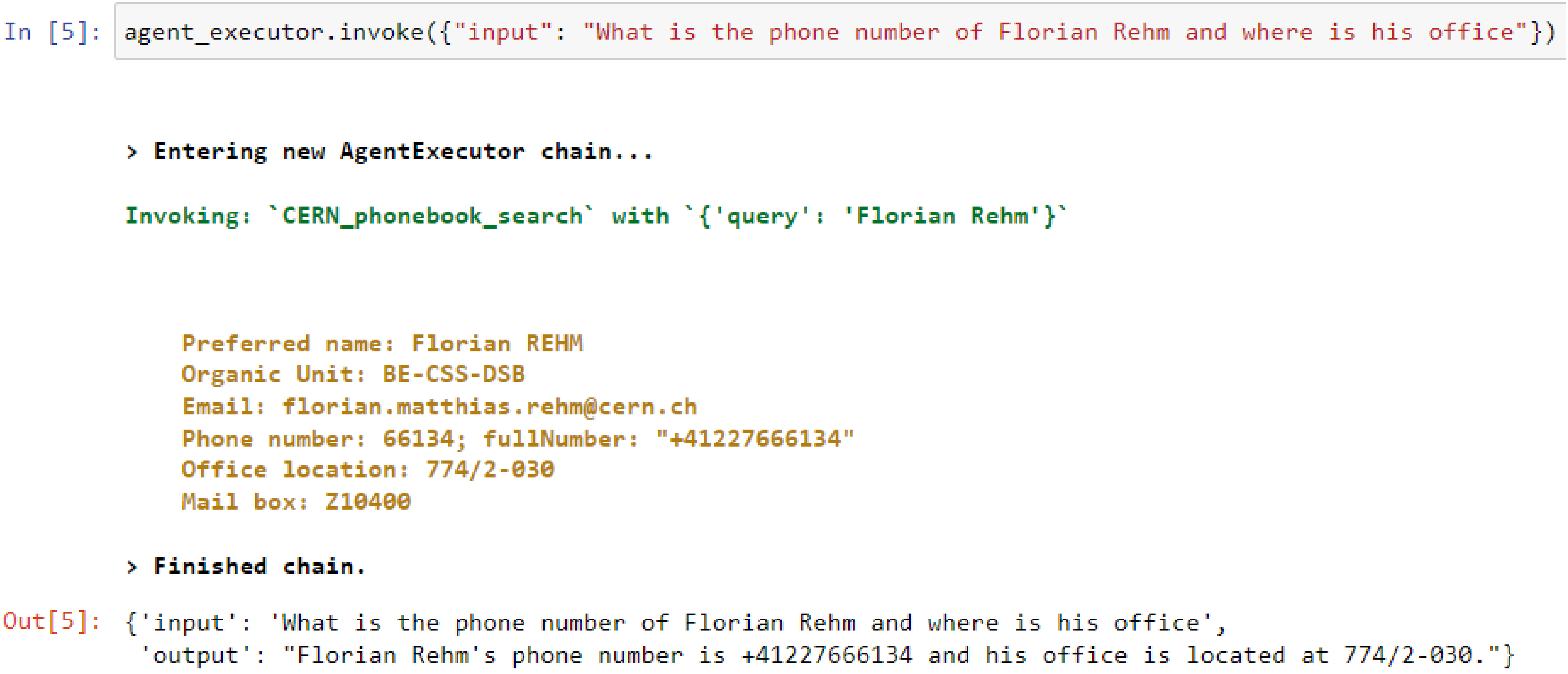
Test with domain-experts:

- Feedback-driven: A community-driven enhancement process for a comprehensive vertical experience.

Further ideas:

- Fine-tune LLM: Customize with CERN specific data to improve performance and relevance.
- Enhancements to coding tasks:
  - Utilize a dedicated coding foundation model.
- Multimodal expansion: Consider plots, pictures, videos (presentations, lectures), …

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:
    This feature will not come (soon),
    due to data privacy reasons.

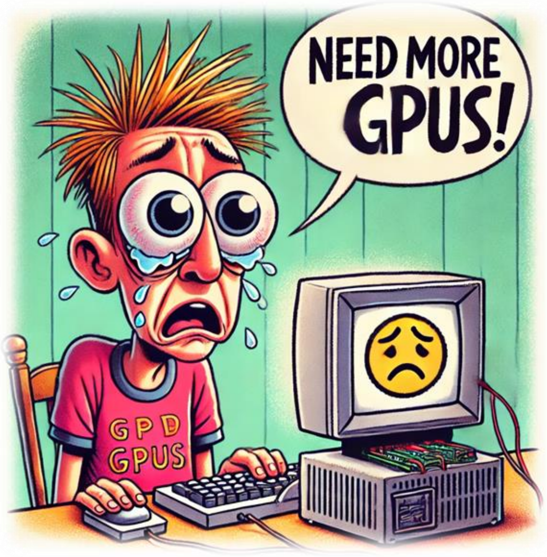
Summary

Summary

- **AccGPT =** a chatbot pilot for CERN specific knowledge retrieval.

- Continuous improvement and knowledge expansion.
  - Already good accuracy. Still a lot to be improved.

- (So far:) Entirely open-source!



By GPT4