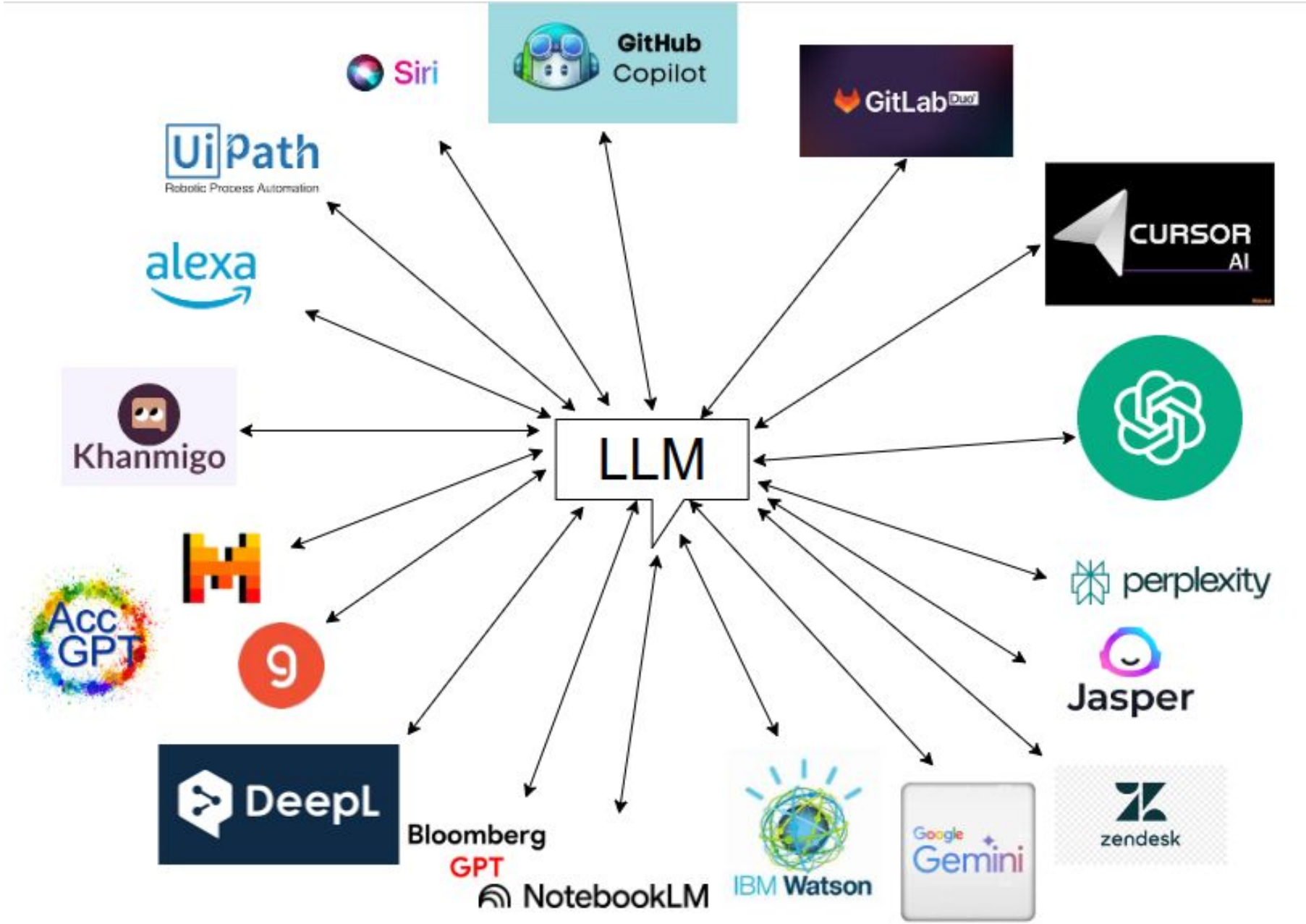
Development of Large Language Models at CERN
(What are LLMs used for?)

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


Siri
GitHub Copilot
GitLab Duo
UiPath
Robotic Process Automation
CURSOR AI
alexa
Khanmigo
LLM
perplexity
Acc GPT
Jasper
DeepL
Bloomberg GPT
NotebookLM
IBM Watson
Google Gemini
zendesk

The LLM Revolution

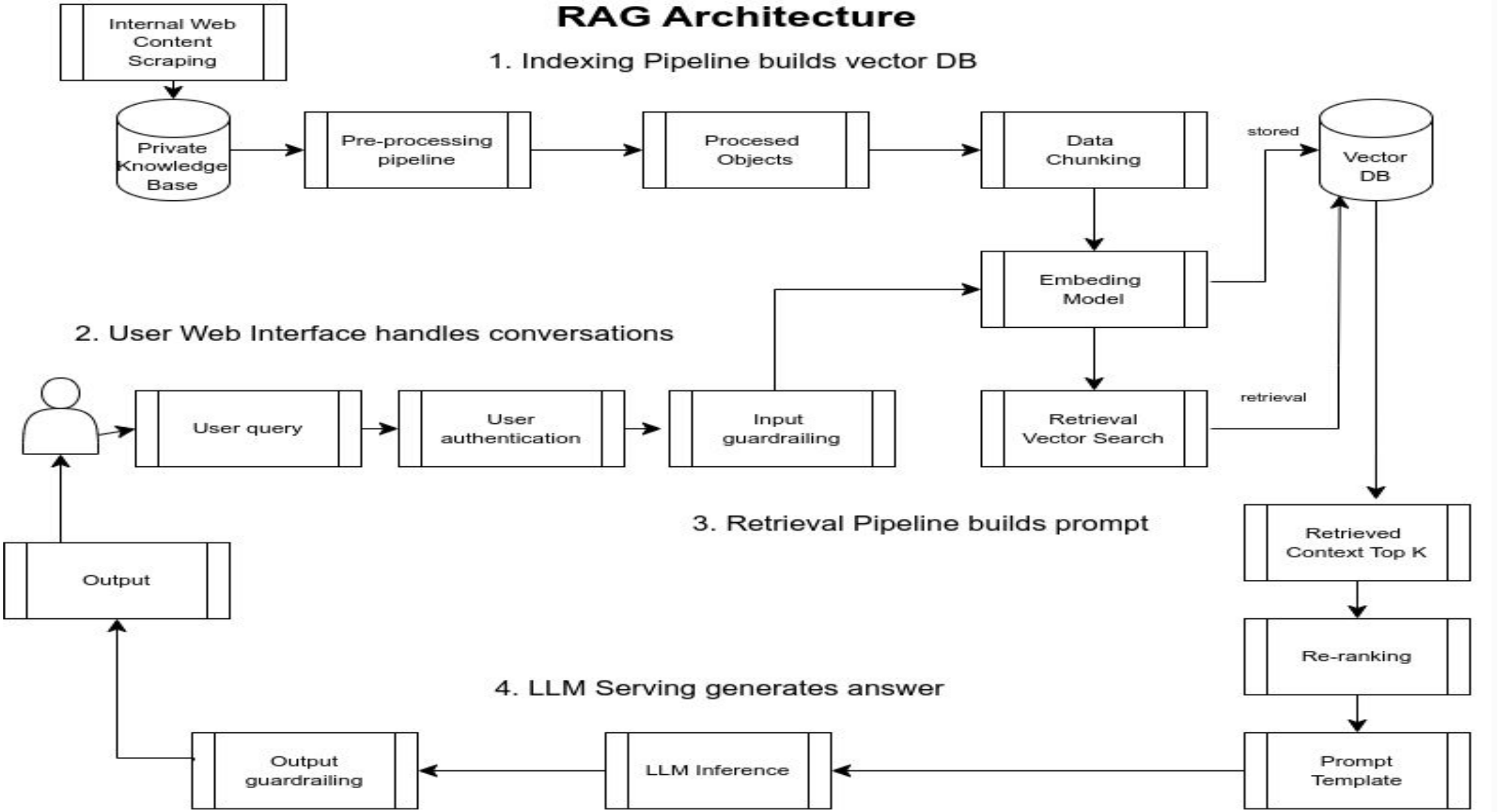
- AI assistants based on **generative AI and LLMs are transformative** across various fields.
- The CERN community is **embracing LLMs to improve tasks** such as information retrieval, code generation, and optimization.
- Efficient, safe, trustworthy, reliable, and reproducible use of AI tools requires **building expertise** and infrastructure at CERN.
- A survey showed that most CERN personnel regularly use AI assistants.
  - **More than 80% of code developers** use code assistants.
- It is CERN's responsibility to ensure these products are available while addressing **data privacy and efficient resource use**

Growing Demand: LLM Use Cases

- The range of applications is broad, spanning specialized scientific content to generic office tasks. Examples include:
  - Knowledge retrieval and code generation.
  - Secure & Personalized Search & Information Extraction.
  - AI-driven Process Automation (AI agents)
  - Speech Recognition & Translation

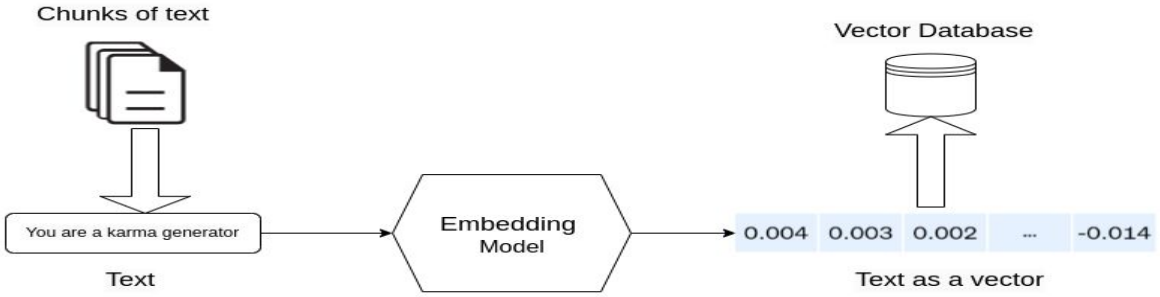
Information Retrieval Use Case

- AccGPT PoC: Knowledge Retrieval
  - Chatbot that leverages Natural Language Processing (NLP) for knowledge retrieval.
  - Can use open-source & commercial LLMs. On-premises & Cloud.
  - Aims at answering questions on CERN internal web content


RAG Architecture
1. Indexing Pipeline builds vector DB
Internal Web Content Scraping
Private Knowledge Base
Pre-processing pipeline
Procesed Objects
Data Chunking
stored
Vector DB
Embeding Model
2. User Web Interface handles conversations
User query
User authentication
Input guardrailing
Retrieval Vector Search
retrieval
3. Retrieval Pipeline builds prompt
Retrieved Context Top K
Re-ranking
Prompt Template
4. LLM Serving generates answer
LLM Inference
Output guardrailing
Output

Retrieval-Augmented Generation

- AccGPT leverages an open-source LLM (Meta-Llama-3.1-8B-Instruct) that is self-hosted & served using vLLM.
- The Indexing Pipeline:
  - Web Content Scraping
  - Preprocessing/Chunking
  - Vectorization/Context Database



- CERN Web content in the generation process.
- RAG retrieves the correct knowledge in >90% of cases and is LLM agnostic.
  - Without RAG: Accuracy drops, as LLM is not aware of CERN
  - RAG (dynamic knowledge) is not fine-tuning (static knowledge)

Lessons learned

- Cloud LLMs: Compliance with **data privacy** and security requirements.
- **Right-size LLMs**: optimize resources and manage costs
  - Quantization / Inference on CPU (testing vLLM on Intel Xeon)
- **User Preferences**: Some users prefer on-site, self-hosted models for continuous operation, confidentiality, and control over data.
- **Varying ACLs** for different content within CERN's internal network
- **Open Source** Community: Embracing open-source solutions can save time and effort. Eg. **OpenWebUI** as user interface framework:
  - Facilitates the development of interactive web applications.
  - Safe Access to LLMs and ways to compare answers / RAG Tools.
  - Multi-modal and versatile, great potential

Use Case: AI-Based Paper Review for CMS

- Collaboration with CERN openlab
  - Student : Annunziata Alvarez-Cascos Hervias (Summer 2024)
  - Initiator : CMS (Chiara Mariotti)
  - Funding : CERN openlab summer student program
  - Supervisor : Florian Rehm
- Problem : Peer review is time-consuming due to:
  - Typos and writing style inconsistencies
  - Non-adherence to guidelines => readability issues
- Solution :
  - Use LLMs to improve paper quality and streamline reviews (8-week proof-of-concept).

Option #1: Fine-tuning

- Fine-tune the Llama 3.1 8B model (Self-supervised, PEFT, LoRA).
- Data: 1,300 peer-reviewed CMS papers (LaTeX source files).
- Results:
  - Took longer than expected (need for more computational resources).
  - Model struggled to understand and correct the scientific content → occasional errors.
- Conclusion:
  - 8B model not powerful enough.
  - **Fine-tuning is complex & resource-intensive.**

```
- Changed the word ``contents'' to ``content''.
- Changed the word ``defined'' to ``defined'' without a space.
- Changed the word ``contents'' to ``content''.
- Changed the word ``defined'' to ``defined'' without a space.
- Changed the word ``contents'' to ``content''.
- Changed the word ``defined'' to ``defined'' without a space.
- Changed the word ``contents'' to ``content''.
- Changed the word ``defined'' to ``defined'' without a space.
- Changed the word ``contents'' to ``content''.
- Changed the word ``defined'' to ``defined'' without a space.
- Changed the word ``contents'' to ``content''.
```

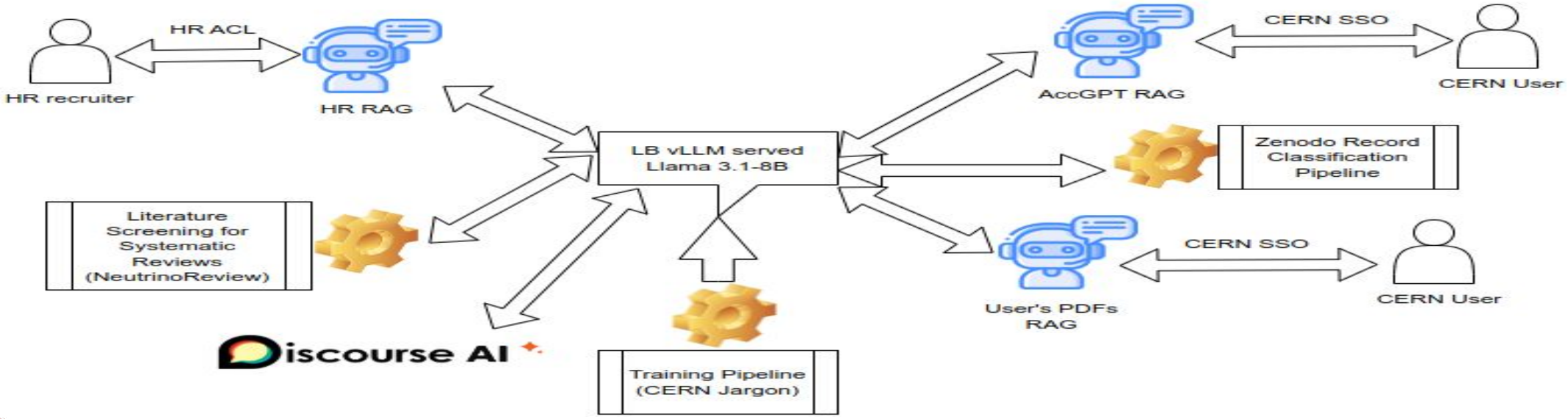
Option #2: Prompt engineering

- Compare raw LaTeX file with the CMS style guidelines: OK
- Project taken over by CMS.
  - Try other fine-tuning techniques / Use more powerful LLMs.

```
Input paragraph:

"The CMS experiment at CERN has recorded a large number of events with a Higgs boson candidate, which is misidentified as a lepton. The events
are selected by a set of selection criteria, including a cut on the transverse energy of the candidate. The transverse energy is calculated using
a fine-grained texture of the calorimeter. The events are then analyzed using a next-to-leading-order calculation which takes into account the
uncertainties of the measurement. The results are presented in a table, which shows the discriminant of the Higgs boson candidate against the background."

Corrected paragraph:

"The CMS experiment at CERN has recorded a large number of events with a Higgs boson candidate, which is misidentified as a lepton. The events
are selected using a set of selection criteria, including a threshold on the transverse energy of the candidate. The transverse energy is calculated
using a fine-grained texture of the calorimeter. The events are then analyzed using a current-current approximation which takes into account
the uncertainties of the measurement. The results are presented in a table, which shows the discriminant of the Higgs boson candidate against the background."

Changes made:

1. "selected by" -> "selected using": Changed the preposition to improve clarity and grammar.
2. "cut on" -> "threshold on": Replaced "cut" with "threshold" to adhere to CMS guidelines.
3. "fine-grained texture" -> "fine-grained texture of the calorimeter": Added a preposition to clarify the meaning and improve grammar.
4. "next-to-leading-order calculation" -> "current-current approximation": Replaced the term with a more formal and accurate description, as per CMS guidelines.
5. "which takes into account the uncertainties of the measurement" -> "which takes into account the uncertainties of the measurement": Removed the unnecessary
phrase to improve clarity and concision.
6. "The results are presented in a table, which shows the discriminant of the Higgs boson candidate against the background" -> "The results are presented in a
table, which shows the discriminant of the Higgs boson candidate against the background": No changes were made, but the sentence is grammatically correct and clear.
```

Need for an LLM Service

- A central LLM service would ensure **efficient resource usage** and can provide ambitious, large-scale solutions.
- It can manage specific use cases, such as offline use in the Technical Network, ensuring **internal data does not leave the CERN network**.
- CERN is considering setting up an LLM service providing **a generic, flexible infrastructure** to easily include additional applications.



Conclusion & Way Forward

- CERN is actively exploring the establishment of a service aimed at:
  - Effectively **indexing** and providing access to CERN-internal information.
  - Offering **secure and reliable access** to both commercial and open-source LLMs.
- CERN IT is currently examining the **foundational components** necessary for CERN user communities to:
  - Implement **RAG pipelines** efficiently.
  - Develop custom **AI agents** tailored to specific research needs.
  - Enhance existing applications by integrating advanced functionalities that **leverage LLM access**.

Questions?

- AccGPT @ CHEP2024
- IT-FTI