

Introduction to Advanced LLM Use Cases

Tutorial: RAG and Agents

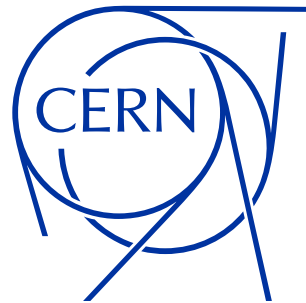
Florian Rehm
11.04.2025

5th ICFA Beam Dynamics Mini-Workshop on Machine Learning for Particle Accelerators at CERN

Agenda

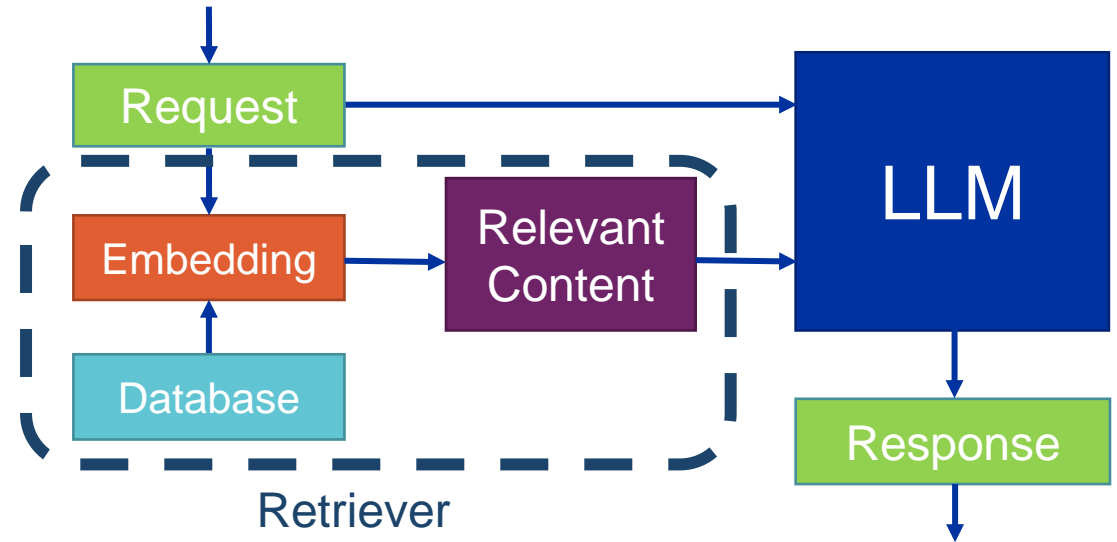
- **Introduction to RAG and Text Retrieval**
- **Live Demo on RAG**
- **AccGPT + Wrap Up**
- **(Short Intro to Agents)**

Text Retrieval with RAG



What is RAG?

- **RAG (Retrieval-Augmented Generation).**
- **Definition:**
 - A hybrid approach that integrates a retrieval system with a language generation model.
- **Key Components:**
 - Retriever: Finds relevant passages/documents.
 - Generator (LLM): Produces answers using both user request and retrieved content.
 - Accompanied by a pre-built data base.



RAG – Key Advantages

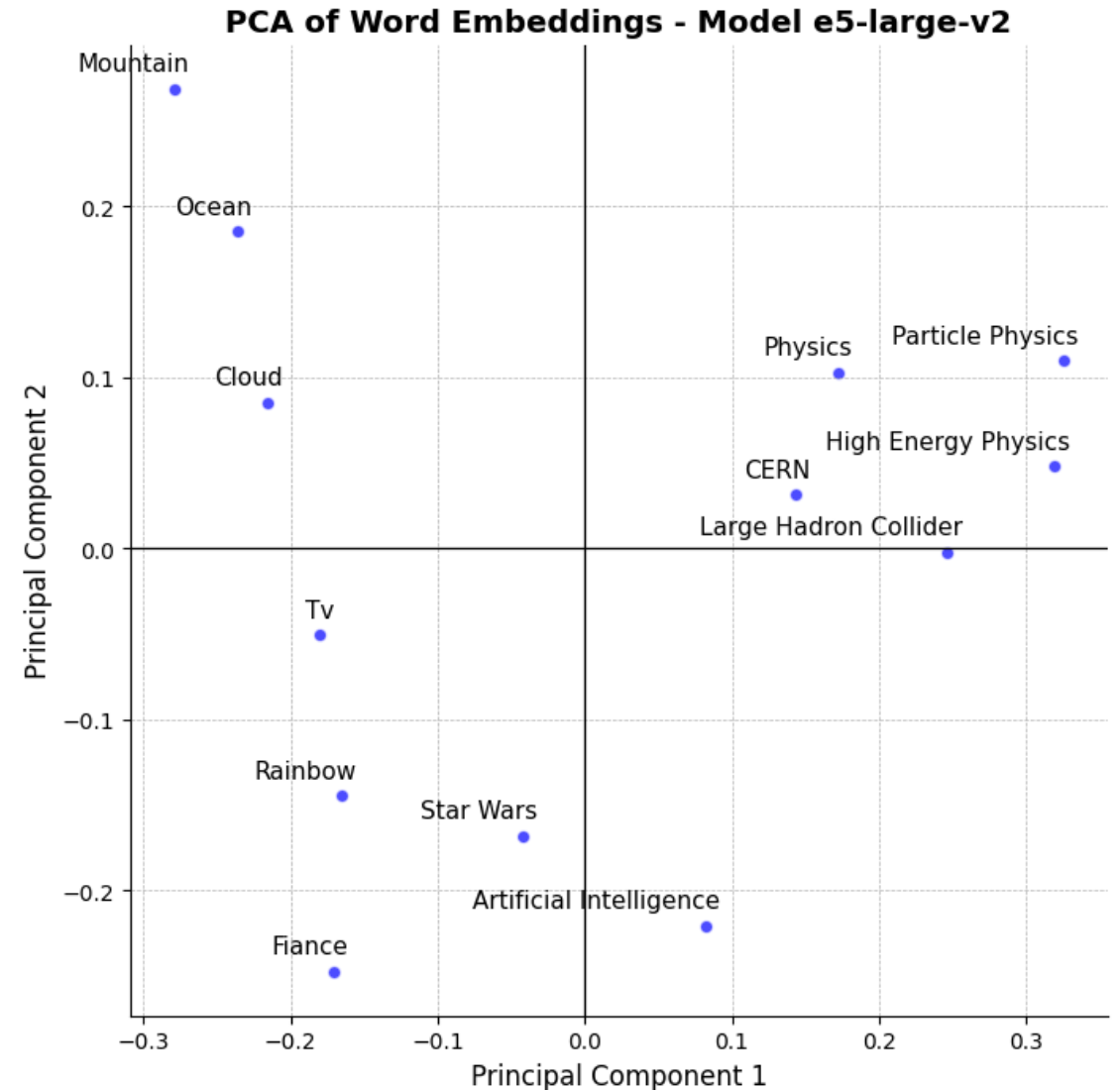
Feature	RAG	Fine-Tuning
Data Updates	Instant via document refresh	Requires retraining
Cost	Lower (no retraining)	High (compute + time)
Hallucinations	Reduced (external grounding)	Higher risk without proper data
Domain Specificity	High (via retrieval corpus)	High (if fine-tuned well)
Scalability	Easier to scale across domains	Less flexible

- **Contextual Accuracy:** Responses grounded in real-time, external sources.
- **Fewer Hallucinations:** Reduces fabricated content by referencing factual data.
- **Up-to-Date Knowledge:** External content can be updated anytime—no retraining needed.
- **Domain Adaptability:** Easily integrates with specific datasets or domains.

Embedding Models

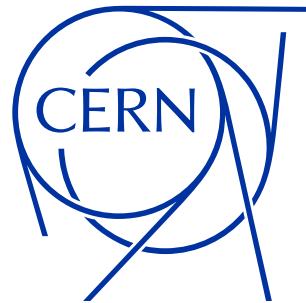
- **Purpose:** Embed high-dimensional input (e.g. text) into low-dimensional, dense vectors.
- **Semantic Mapping:** Similar inputs have similar vector representations (due to training).
- **Use Cases:** Similarity searches between embedded vectors.
- **Architecture:** Based on transformers.

- **Example: Principal Component Analysis (PCA).** Dimensionality reduction technique that transforms correlated variables into a smaller set of uncorrelated variables called principal components, preserving as much variance as possible.



RAG in Practice

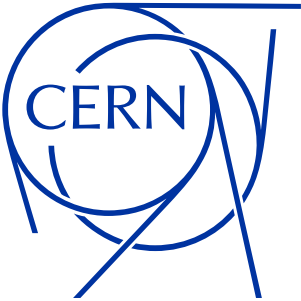
Coding!



RAG Challenges

- **Retriever Quality:**
 - Depends strongly on data quality, data preprocessing (chunking), embedding model, ...
- **Latency:**
 - Added retrieval step may increase response time.
- **Context Integration:**
 - Balancing prompt length and relevance (how many retrieved text chunks to include into the LLM?).
- **Scalability:**
 - Efficient handling of large document collections.
 - Overlapping documentations.

Example use case: AccGPT



Example Use-Case: What is AccGPT?

- **An advanced RAG chatbot.**
- **Goal: a chat-based search tool for CERN internal knowledge.**
- **In addition to what was shown in the tutorial, AccGPT can:**
 - Query rewriting by LLM + key word / key phrases extraction by LLM.
 - Multiple similarity searches: on question + key words / key phrases.
 - Takes previous and subsequent text chunks to the one retrieved, if same domain (for longer documents).
 - Using meta data in text chunks (advanced data preparation).
 - Including multiple relevant text chunks for answer generation (from different sources).
 - Applying a re-ranker (cross-encoder) model on retrieved text chunks.
 - Utilizing a vector database to return further chunk information (URL).
 - And more ...

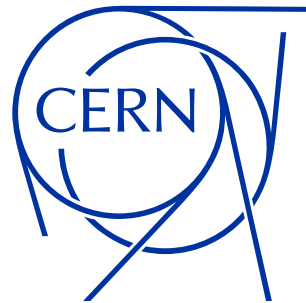
What does it look like?

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, containing the question 'What are the supported Linux distributions at CERN?' and the answer 'Linux at CERN'.

What does it look like?



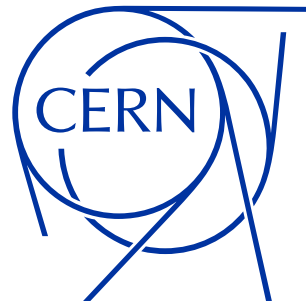
RAG: Key Points



RAG: Key Points

- **Accuracy depends on: Data preparation, data quality and retrieval pipeline, ...**
- **Advantages of RAG vs. fine-tuning:**
 - Easy and fast to update context (DB update, add new context, ...).
 - “Simpler” and “cheaper” to implement / test.
- **Disadvantages:**
 - Model might lack deeper domain knowledge.
 - RAG provided information maybe incomplete.
- **Ideal solution:**
 - Combine RAG with fine-tuning!?

Extra: Agents

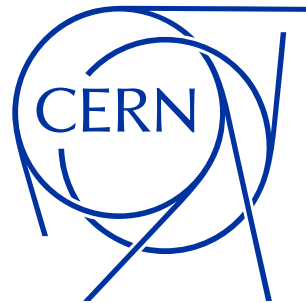


What are LLM Agents?

- **Definition:**
 - AI-driven systems leveraging LLMs to autonomously perform tasks and interact with tools.
 - **Tools:** External modules or APIs enabling agents to perform tasks beyond text generation (e.g., databases, code execution, web search, ...).
- **Key idea:**
 - Combine reasoning, planning, and execution.
- **Use cases and impact:**
 - Automating complex workflows (research, data analysis).
- **How do they work?**
 - User gives a high-level instruction.
 - Agent interprets, plans steps, selects tools, executes actions.
 - Example Tool: Python environment – enables an LLM agent to run code, analyze data, or visualize results.
 - Iterative process: continuously refining based on intermediate results.

Agents in Practice

Coding!



Example Use-Case: What is AccGPT-Pro?

- **Future advanced model, currently under development.**
- **Based on Agents.**
 - One tool (the key tool!) is AccGPT.
 - Could be augmented with keyword search, knowledge graphs, ...
- **Agentic workflow:**
 - The Agent plans and executes the retrieval.
 - Starts with a planning step: “For what context should it search for?”
 - Applying one or multiple AccGPT context searches.
 - Intermediate planning: “Is already all the required context retrieved?”; “Based on the retrieved results, is there more information required?”
 - Potentially do more AccGPT searches.
 - Provide and phrase the final response.
- **Potential to provide much better and more grounded responses than AccGPT.**
- **Downside: much more computationally demanding.**

That's it!

