
CERN

ItGPT:
an AI-Powered Assistant and more…

M. Guijarro IT-CD-CLI

Why CERN needs an AI chatbot? I

- Growing demand for a comprehensive chatbot service
  - **BE-ICS-FT** for automatic code generation and translation
  - Integration with video and transcription services (**TTaaS**)
    - Generating notes, summaries, whiteboards, emails, etc
    - Integration/generation of data sets
  - General questions on programming, computing or science
- ChatGPT (and Github Copilot) professional usage at CERN
  - Need a central OpenAI subscription for the whole of CERN
  - Need to understand different use cases
- Expectations for support have changed (MM before tickets)
  - Users prompt questions to search engines

Why CERN needs an AI chatbot? II

- Concerns with the free version of ChatGPT
  - Free version of ChatGPT has **limited features**
  - ChatGPT is **unacquainted with CERN internal data**
  - ChatGPT could generate misinformation:
    - The data set it works off of is up until 2021
    - Hallucinations from data (large sets) or training
  - ChatGPT could be used for malicious purposes
  - Importance of **data privacy** when using AI tools
  - Understanding privacy policies and terms of service
  - Risk of leakage of CERN confidential data: Samsung case

Proposal: Initial Implementation (p1)

- Explore creation of an ItGPT-web portal for user interaction
  - Azure OpenAI API for backend processing
  - Data Privacy: **European end-point** for alignment with CERN policies
- Channel OpenAI API usage at CERN:
  - Further **understand our use cases**
  - **Enrich answers** with CERN specific knowledge: LangChain
  - Querying different specialised LLMs
  - Avoid **vendor lock in**: Away from OpenAI LLM
  - Gather/Parse Q/A to derive training datasets
- Initial use case: **Self-support** and answering IT service questions
- Pilot could be ready within 3 months

LangChain flow

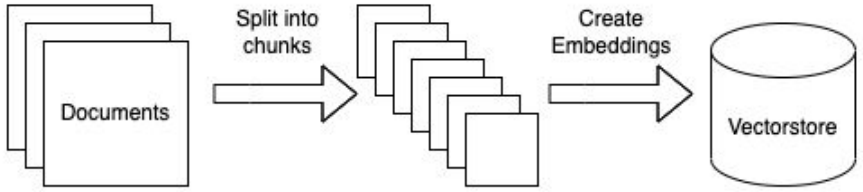


Diagram of typical ingestion process

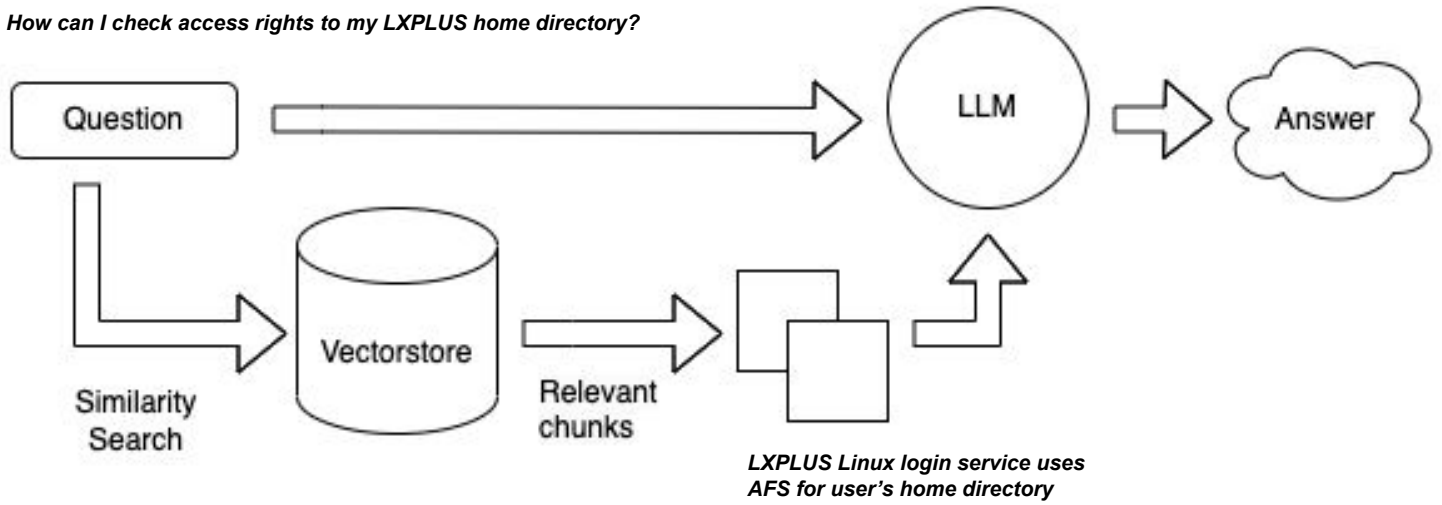


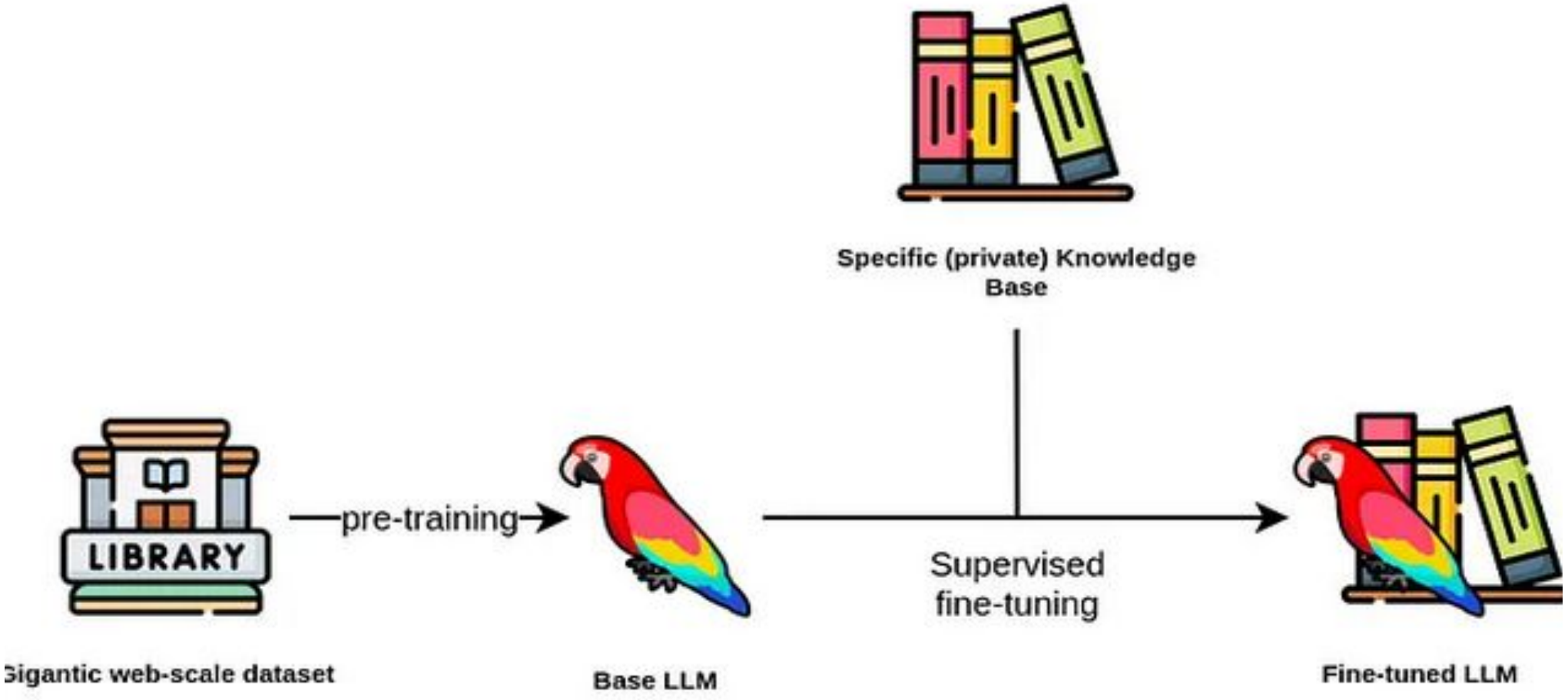
Diagram of typical query process

Source: https://blog.langchain.dev

Proposal: Project Evolution (p2)

- Evaluate Open Source LLMs as alternatives
  - Easy setup, open-source nature, diverse data collection, cost-effectiveness, and promising functionality
- Training with CERN Internal Documentation/**Code**
  - **Enhanced accuracy**, **improved security**, customized knowledge base, and increased productivity
  - **Knowledge and CI/CD pipelines sharing**, improved language model, enhanced reputation, and collaborations
  - Identifying **how to structure our documentation** so that an LLM could be trained on this data in the future
  - Better **understand CERN (old) codebase**

LLM Supervised fine-tuning Pipeline



Source: medium.com. Icons from Flaticons

Expected Impact

- Enhanced Productivity (extension of CERN Search engine):
  - Personalized assistance and **easy access to relevant information**.
  - Advanced search and recommendation systems enable **faster information retrieval**.
- **Reduce cost of providing User Support** (0.25 grade-8-FTE)
  - Support teams spend less time on repetitive questions
- Support not limited to working hours + faster reply (24/7)
  - CERN users get answers faster
  - Offer **multilingual customer support**
- Competitive Advantage:
  - Cutting-edge AI implementation demonstrates **technological leadership**
- Faster evolution of **CERN codebase**
  - Unit tests production, code translation, faster debugging of old code

Innovation factors

- Gain experience on Artificial Neural Networks for Natural Language Processing (NLP) and **Generative AI,** which the new http protocol
- Establish a service which consumes the facilities and resources IT Dept provides for ML activities: **Eat our own food**
- Improve current **documentation processes** which are time-consuming, manual, and prone to errors
- Address difficulty in **locating relevant information**, which hampers productivity and efficiency
- Provide interactivity and personalized assistance to **improve user experience**
- **Optimize resource** usage: Run models in 8-bit or less

ItGPT Project proposal

- 1.4 FTEs: 0.4 Staff member + 1 Fellow
- Web app interfacing with OpenAI API ~ 3 months
- Open Source LLMs evaluation + Training with our doc ~ 1 year
- IT Dept resources only:
  - Access to OpenAI API: $0.002 per 1,000 tokens (**p1**)
  - Access to (8) GPUs on servers with 16 to 512GB RAM and up to a few Terabytes of Storage (**p2**)
  - Applying to the 2023-24 EU calls on AI topics, such as AI4EU projects, could be considered
- Status:
  - Tried OpenAI API + GPT4All
  - Comparative papers on LLMs (Falcon) + background research

Conclusion

- ItGPT, an AI-Powered Assistant project offers a transformative solution to streamline documentation processes, **improve user experience, and enhance productivity**. By leveraging AI technology, CERN can revolutionize its documentation practices, **leading to increased efficiency and accuracy**
- Its alignment with EU-funded AI4EU initiative opens **doors for external funding and collaboration**. This opens doors to additional resources and expertise, further enriching the project's potential impact
- Approving this project will empower CERN with **advanced AI capabilities and a competitive edge (and collaboration) within HEP sites**. It signifies CERN's commitment to innovation, efficiency, and **technological leadership**
- Our record shows **we can**: Quattor->Puppet; LSF->HTCondor; ES->OS

Beyond the scope of this project…

- An **unstoppable force**:
  - The use of OpenAI LLM will transform MS tool suite
  - Similarly, Google's LLMs will transform their tools suite

- What will be the role of the IT department on transforming CERN tool suite?:
  - ROOT, GEANT, AliRoot, CMSSW, CERNVM, MadGraph, LHCb Software Framework, Gaudi, FAstJet, CAP, and more
  - LLMaaS: **Custom LLMs** to drive innovation and address **research-specific challenges**
  - We'll need a Generative AI group in the IT Dept

- The biggest risk is CERN (IT Dept) **not taking action about this**:
  - Uncontrolled use of commercial LLMs (**$$$**) + Data privacy concerns
  - Outsourcing CERN's core activity as **Knowledge Source**
  - Missing one more chance for **Digital sovereignty** for Europe
  - **A chance to bite the bullet, which is anyway going to hit us**

Questions?

Work ahead of us

- Chatbot Backend Selection/Evaluation:
  - OpenAPI and LangChain integration
- Chatbot Tool Development and Testing:
  - Web Development and Chatbot Integration
- Open Source LLM Selection/Evaluation
- LLM Training
- Automation of Documentation Feeding:
  - **Implement automation Tools and CI/CD Pipelines**
- Monitoring and Evaluation: Final Backend Architecture
- Scaling up to Other Sets (Documentation/Code)
- Skills needed DevOps, Data Preprocessing Techniques, NLP and Deep Learning Algorithms
- Eg. Falcon Training cost around 2700 petaFLOP-days, 75% that of GPT-3