

Multi-modal LLMs at EuXFEL

Knowledge Assistants, Data Exploration, and Policy



L. Brodersen, I. Derevianko, D. E. Ferreira De Lima, L. Gelisio, S. Hauf, M. Manetti, R. Rosca, F. Sohn, K. Wrona, et al.

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

CERN, Friday, 11.04.2025

Contact: steffen.hauf@xfel.eu

Introduction

- Use cases and policy drafts in this presentation are the outcome of a task force to suggest a policy for general LLM usage at the European XFEL
- **2024: Charge: Evaluate the opportunities and risks associated with using AI Large Language Models at EuXFEL over a six months period and provide a report to EuXFEL Management.**
 - Evaluate legal and technical implications
 - Evaluate use cases
- **2025 - I: Develop a policy for the facility, and technical means for generic access**

Artificial Intelligence: opportunities and risks of using Large Language Models (ChatGPT)

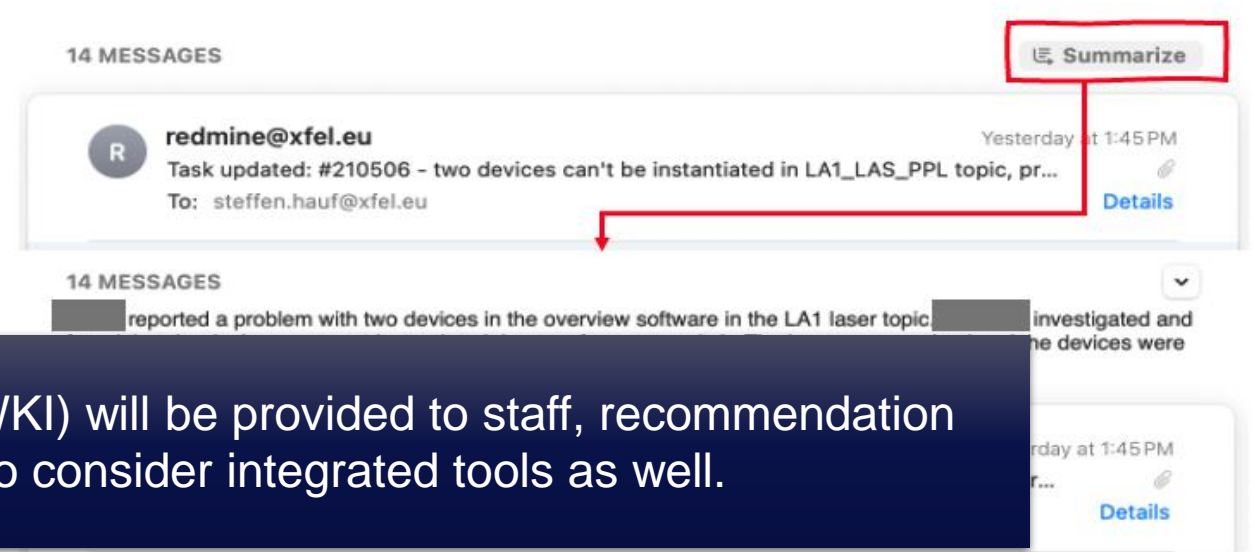
Lisa Brodersen,
Steffen Hauf



European XFEL

Text summarization, Editing & Text Composition

- Likely most prominent use case of LLMs in the public perception
- Current state-of-the-art models generally very capable, but
 - What about ...
 - What about ...
- Where we are ...
 - Integrated capabilities in Windows and Mac OS
 - More on-device/edge processing

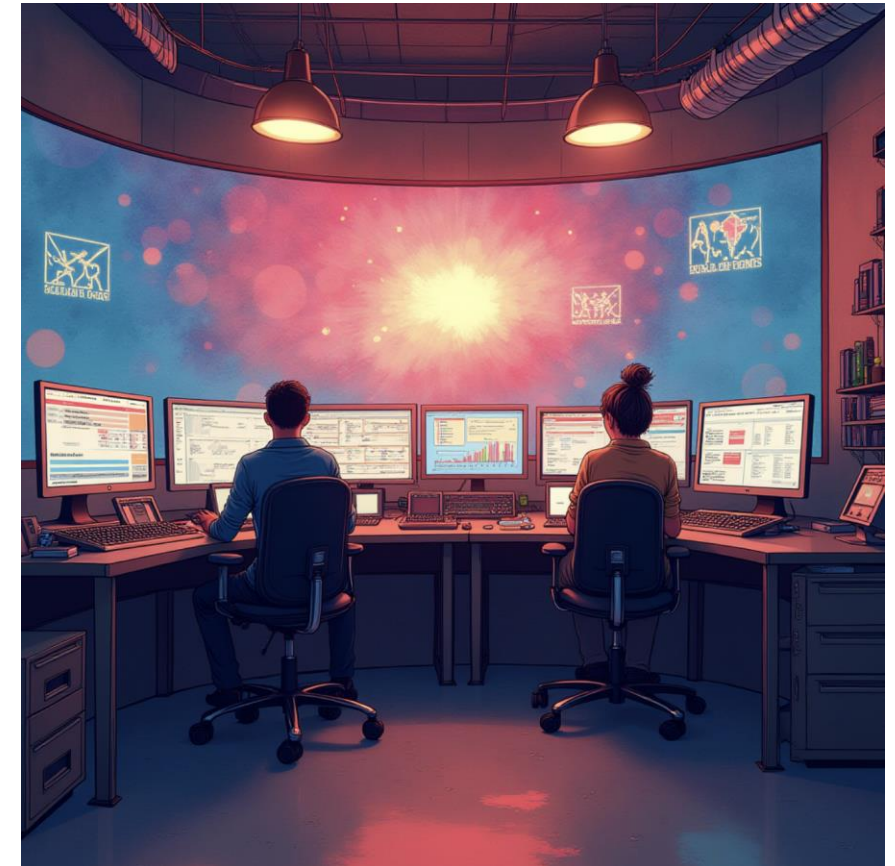


While a generic interface (HAWKI) will be provided to staff, recommendation to EuXFEL management was to consider integrated tools as well.

An email thread of 14 messages summarized by Apple Intelligence as integrated in the Mac OS Sequoia Beta

DOCbot Knowledge Assistant

- Staff from various Data Department teams rotate shifts in the Data Operation Center (DOC) to monitor and support Data services, with 2nd level on-call support available to handle complex issues.
- Resolved issues are documented in Redmine tickets, creating a knowledge base over time that includes problem descriptions and solutions.
- A knowledge assistant bot, integrated into the Zulip chat tool, uses tickets information (Retrieval Augmented Generation, RAG) to assist in issues classification and resolution through text-based interaction.



Comic style light warm colors watercolor - two people free electron laser brightly lit friendly atmosphere control room many screens control software normal desks

DOCbot Knowledge Assistant

- An experiment involving Data staff was conducted to assess the practical applicability of such knowledge assistant.
- Staff members were invited to participate in a survey and interact with the knowledge assistant to resolve one of three problems, each varying in clarity, domain relevance, and level of detail.
- The experiment provided an opportunity to compare a state of the art commercial service (**GPT-4o**, bot Alice) against a locally hosted open source LLM (**Vicuna**, bot Bob): experiment participants interacted with one of the two bots.

DOC_BOT_Experiment > ✓ Test Bot #42

AI Team
@Alice Trying to move a motor gives an error; apparently the motor does not move

Alice
There are multiple issues related to motors not moving and giving errors. Here are the relevant problem summaries and their Issue_IDs:

1. Issue_ID: 128103
 - o Problem: MID MID_EXP_UPP/MOTOR/T16 does not move. In Karabo, the motor appears healthy but it fails to move. After several attempts to issue moving commands, the motor encounters an error.

[SHOW MORE](#)

I found 31 related issues from the knowledge base. You can ask me questions about these issues, or send 'clear' to start a new search.

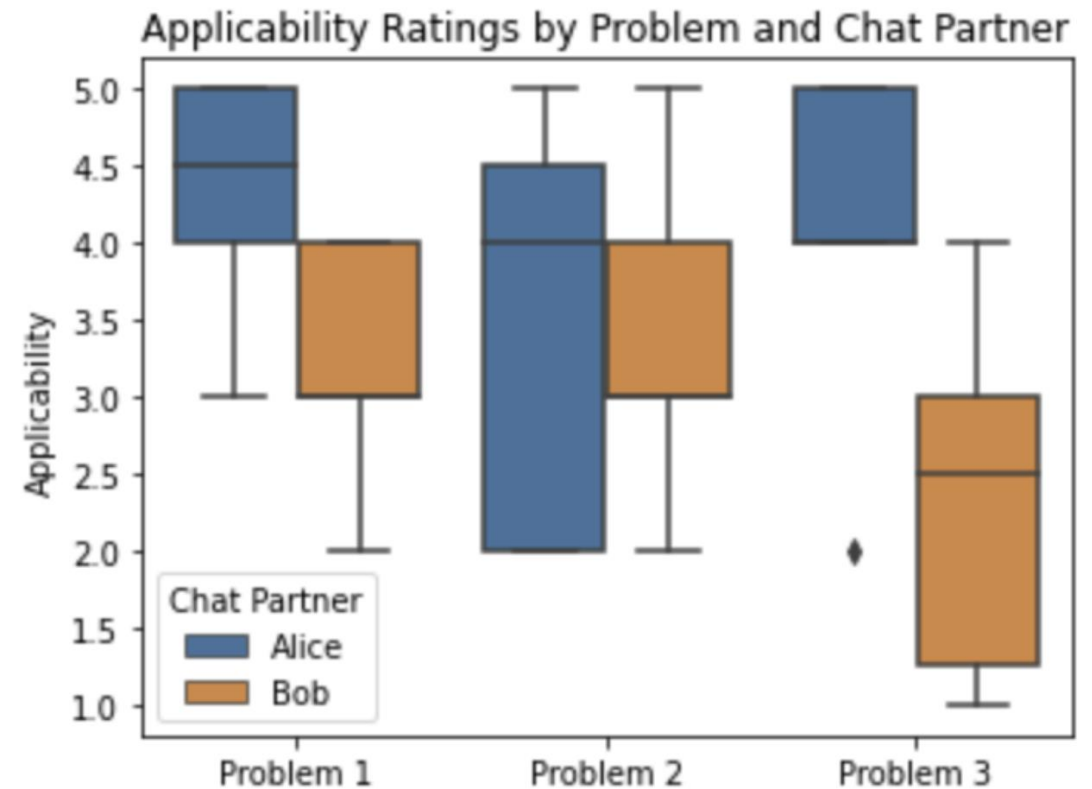
AI Team
@Alice how many issues are related to hardware versus software problems

Alice
Based on the provided context, here is a breakdown of issues related to hardware versus software problems:

Interaction with bot Alice (linked to a state of the art commercial LLM) via the Zulip chat tool

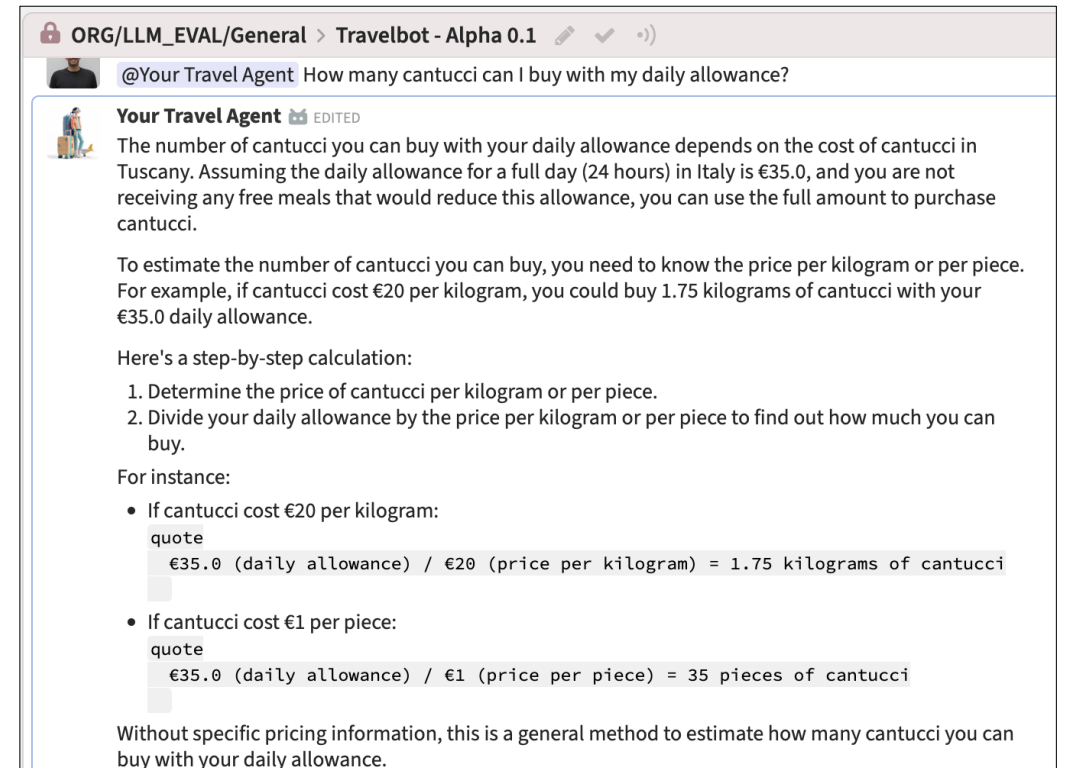
DOCbot Knowledge Assistant

- **Perceived Value:** AI knowledge assistants are viewed as beneficial for technical tasks and managing semi-structured information, making their development worthwhile.
- **Trust & Caution:** Users tend to overtrust AI responses, pointing out the need for usage policies that limit AI to low-risk domains, where errors have minimal impact.
- **LLM Comparison:** The commercial GPT-4o performs better than the local, open-source Vicuna for generic or lower-quality inputs, though users noted a smaller performance gap than expected.




Travelbot Knowledge Assistant

- Staff participating in this experiment utilized a RAG Chatbot to plan and file expenses for a business trip.
- The prompt was crafted to enable the Language Model (LLM) to access general knowledge on travel regulations and planning.
- The bot's primary knowledge base was the EuXFEL travel guidelines.
- Two bots, Adira and Paul, were implemented. Adira interfaced with the European **Mistral Large** model from Mistral.AI, while Paul interfaced with OpenAI's **GPT-4o** model.



ORG/LLM_EVAL/General > Travelbot - Alpha 0.1

@Your Travel Agent How many cantucci can I buy with my daily allowance?

Your Travel Agent  EDITED

The number of cantucci you can buy with your daily allowance depends on the cost of cantucci in Tuscany. Assuming the daily allowance for a full day (24 hours) in Italy is €35.0, and you are not receiving any free meals that would reduce this allowance, you can use the full amount to purchase cantucci.

To estimate the number of cantucci you can buy, you need to know the price per kilogram or per piece. For example, if cantucci cost €20 per kilogram, you could buy 1.75 kilograms of cantucci with your €35.0 daily allowance.

Here's a step-by-step calculation:

1. Determine the price of cantucci per kilogram or per piece.
2. Divide your daily allowance by the price per kilogram or per piece to find out how much you can buy.

For instance:

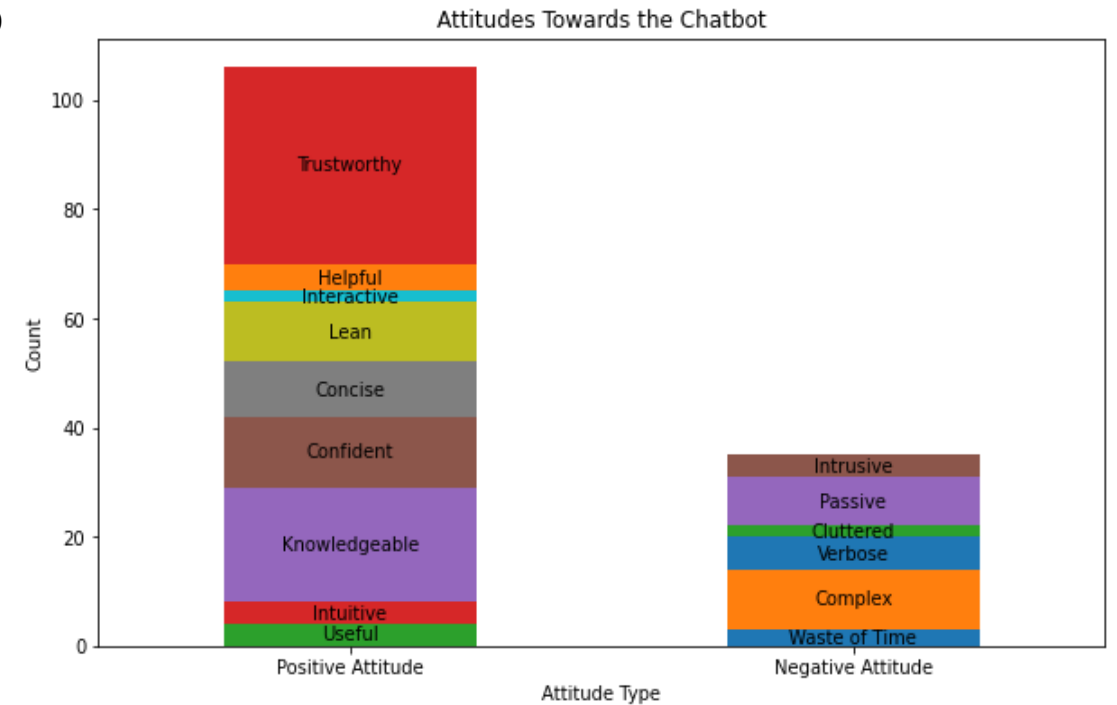
- If cantucci cost €20 per kilogram:
quote
$$\text{€35.0 (daily allowance)} / \text{€20 (price per kilogram)} = 1.75 \text{ kilograms of cantucci}$$
- If cantucci cost €1 per piece:
quote
$$\text{€35.0 (daily allowance)} / \text{€1 (price per piece)} = 35 \text{ pieces of cantucci}$$

Without specific pricing information, this is a general method to estimate how many cantucci you can buy with your daily allowance.

Interaction with a first prototype Travelbot (linked to OpenAI's GPT-4o) via the Zulip chat tool: the bot can be queried about topics beyond the EuXFEL travel guidelines.

Travelbot Knowledge Assistant

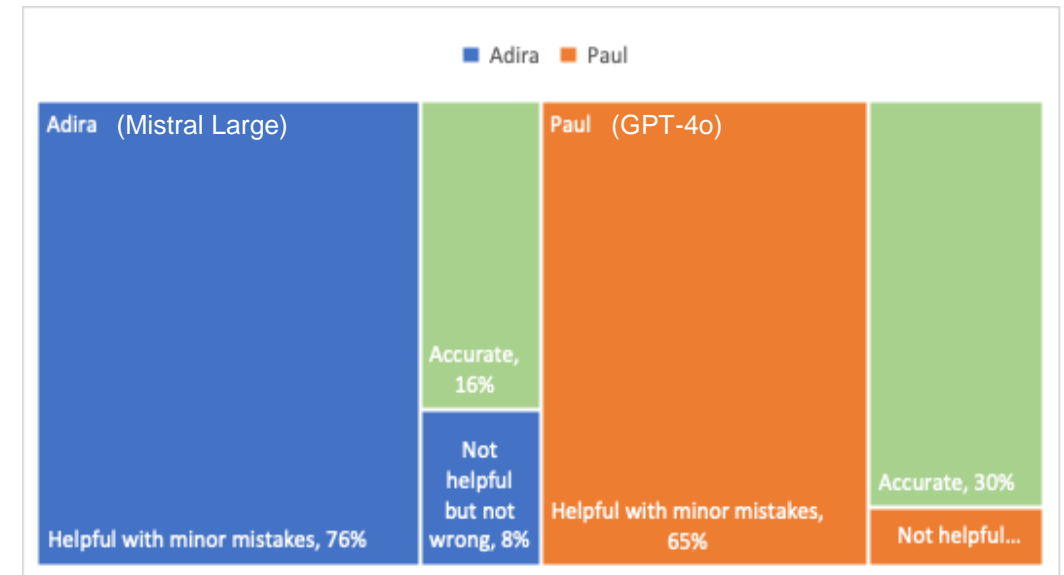
- Experiment participants were presented with one of two scenarios:
 - planning a business trip to a conference in Grenoble, which included a private stay in the city
 - filing travel expenses for this trip.
- Participants were required to provide demographic information and feedback on the AI tools utilized.
- Staff members who have undertaken at least one business trip for the facility tend to utilize the chatbot responses for their answers more frequently.
- Overall, participants expressed a positive attitude, as shown in the picture.



Attributes associated to the chatbot by participants

Travelbot Knowledge Assistant

- Travel office staff categorized the bot answers in the chat log in order to assess the accuracy of the bot responses in relation to actual regulations.
- None of the responses from the bots were rated as extremely good or problematic.
- Based on the GPT-4o model, Paul's responses exhibited an average of twice as many accurate answers compared to Adira's responses generated by Mistral.AI.



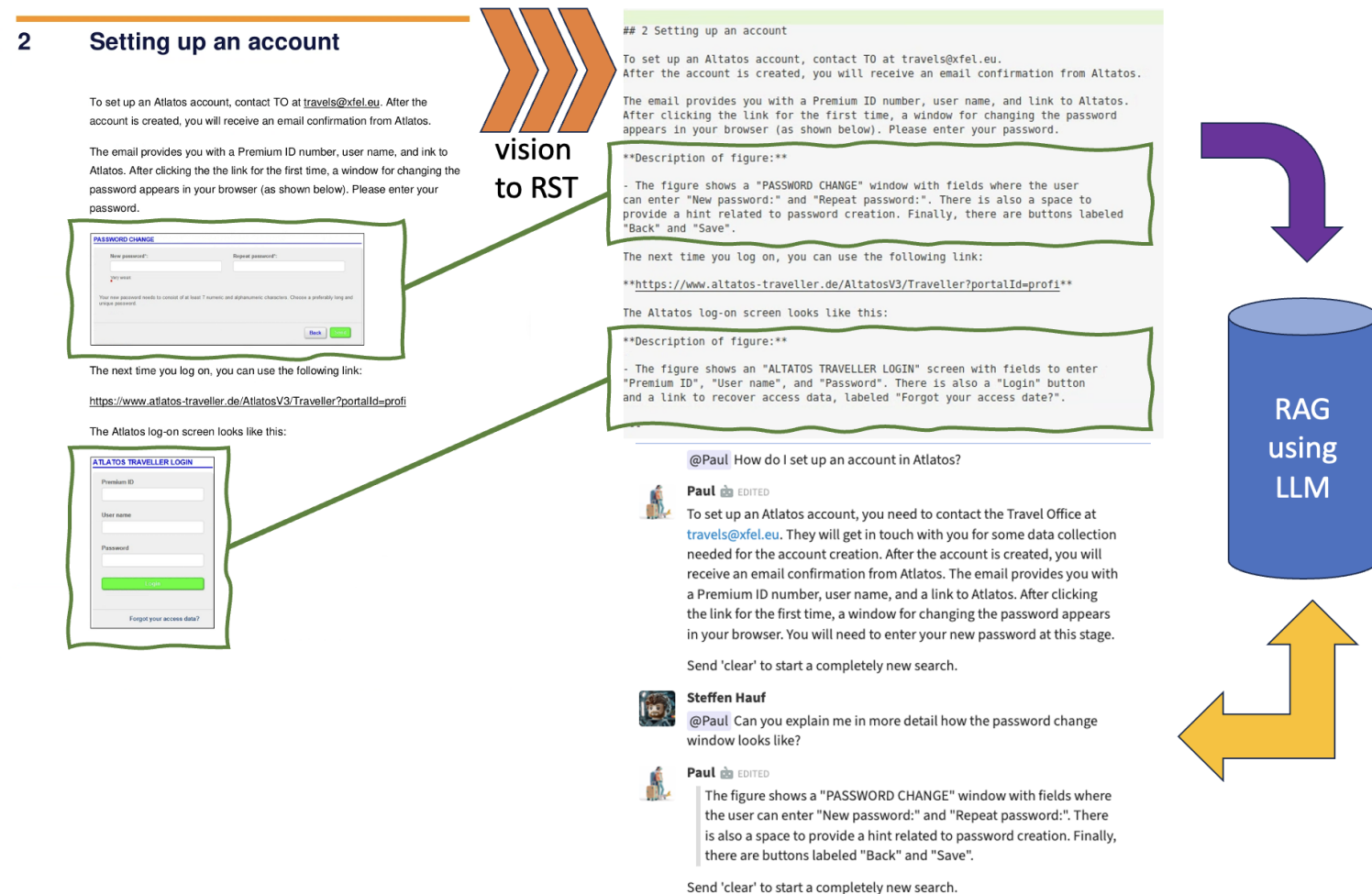
Answer classification by the travel office for the two models.

Adira: Mistral.AI

Paul: GPT-4o

Information Exploration

- A lot of information at EuXFEL is in form of PDFs and Word documents
- While practical, these are opaque information stores
 - Cannot efficiently search across many documents
 - Content is usually not very structured
- Vision-enabled LLMs are very efficient in converting this data into structured data that is **F**indable, **A**ccessible, **I**nteroperable, **R**eusable



General Access: HAWKI

- Tool used by many universities in Germany to provide generic access to generative AI systems. Key priorities include social accessibility, data privacy, and adapting AI tools to academic needs.
- Initially developed at the Hildesheim University of Applied Sciences, HAWKI is an open source web interface allowing to access GPT models via OpenAI's API with no link to personal accounts
- HAWKI acts as a proxy to any requests to the providers, such that it is relatively straight-forward to change to the services of a provider other than OpenAI should a need arise.

The screenshot displays the HAWKI web interface. On the left is a sidebar menu with the following items: Conversation (with a help icon), Chat, Virtual Office (with a help icon), Team, Learning Space (with a help icon), Scientific Work (with a double arrow icon), Data Analysis (highlighted with a blue underline), Feedback, Methodology, Literature Search, Research Support, Writing Help, Organization, and Creativity. The main content area shows a chat conversation. A user message (MM) asks for help with data analysis. An AI response (AI) provides a general approach and lists two main steps: 1. Understand Your Data and 2. Data Cleaning and Preparation. Below the AI response is a text input field with a placeholder 'You can make your request here...' and a blue arrow icon. At the bottom right of the chat area, it says 'Version 1.0 - Model GPT4-Turbo'.

HAWKI

Conversation ⓘ
 Chat
 Virtual Office ⓘ
 Team
 Learning Space ⓘ
 <> Scientific Work

Data Analysis
 Feedback
 Methodology
 Literature Search
 Research Support
 Writing Help
 Organization
 Creativity

MM I have recently conducted an extensive data collection and am now looking for support in analyzing the data. The data collection includes [description of the data], and I would like to [goal of the analysis]. Could you please assist me in analyzing the data and provide advice on interpreting the results?

AI Certainly, I'd be glad to help guide you through analyzing your data and interpreting the results. Since the specifics about your data collection and analysis goals haven't been provided, I'll outline a general approach you can tailor to your needs. If you can share more details about your data and analysis goals, I could offer more targeted advice.

1. Understand Your Data

- **Data Type:** Identify if your data is quantitative (numerical) or qualitative (non-numerical, such as categories or descriptions).
- **Data Structure:** Note how the data is organized—into categories, across time periods, by geographic location, etc.
- **Data Source and Quality:** Understand where your data comes from and assess its reliability and accuracy.

2. Data Cleaning and Preparation

- **Missing Data:** Decide how you will handle missing values (e.g.,

ⓘ
 You can make your request here... ⓘ

Version 1.0 - Model GPT4-Turbo

Specific Applications requiring API Access

- Most of the presented use cases use the service-supplier API
 - Tailored solutions & batch processing
 - Integration with existing services

- Suggestion: request form for such access to be granted
 - Technical vetting by DATA
 - Regulatory vetting by Legal
 - EuXFEL management pre-approves common use cases for non-sensitive data

- Estimate token-costs, and if high contribute to LLM budget of facility

Annex 1 to Staff Note: Request for Foundational Model API Access

Please send the filled and signed form to ai@xfel.eu.

Requested by Group

Name of Application

Brief description of the application:

Target Audience (who will be using it?)

Does your application process any of the following types of data:

Personal Data Images of people Confidential Data ~~EuXFEL~~ Scientific Data

Copy-righted data outside ~~EuXFEL's~~ copyright Behavioral Data of Persons

Does your application fall into any of these pre-approved categories:

AI Knowledge Assistants using only non-sensitive data. If you checked any of the data type fields above, this does likely not apply.

AI Co-pilots using only non-sensitive data. If you checked any of the data type fields above, this

Suggested Implementation – Local Hosting

- Small models are becoming more capable especially for constrained tasks
 - Google Gemma 1.5
 - Microsoft Phi 3.5
 - Flux.Dev & Flux.Schnell
- Plan is to maintain a GPU node hosting the most capable models possible for specific applications to test against



A 8-bit game view of the physicist James Maxwell sitting on top of number of servers, with the Maxwell equations in the foreground

Outlook

- Working deadline for policy and generic access tools is May 2025
- Specific applications are expanding

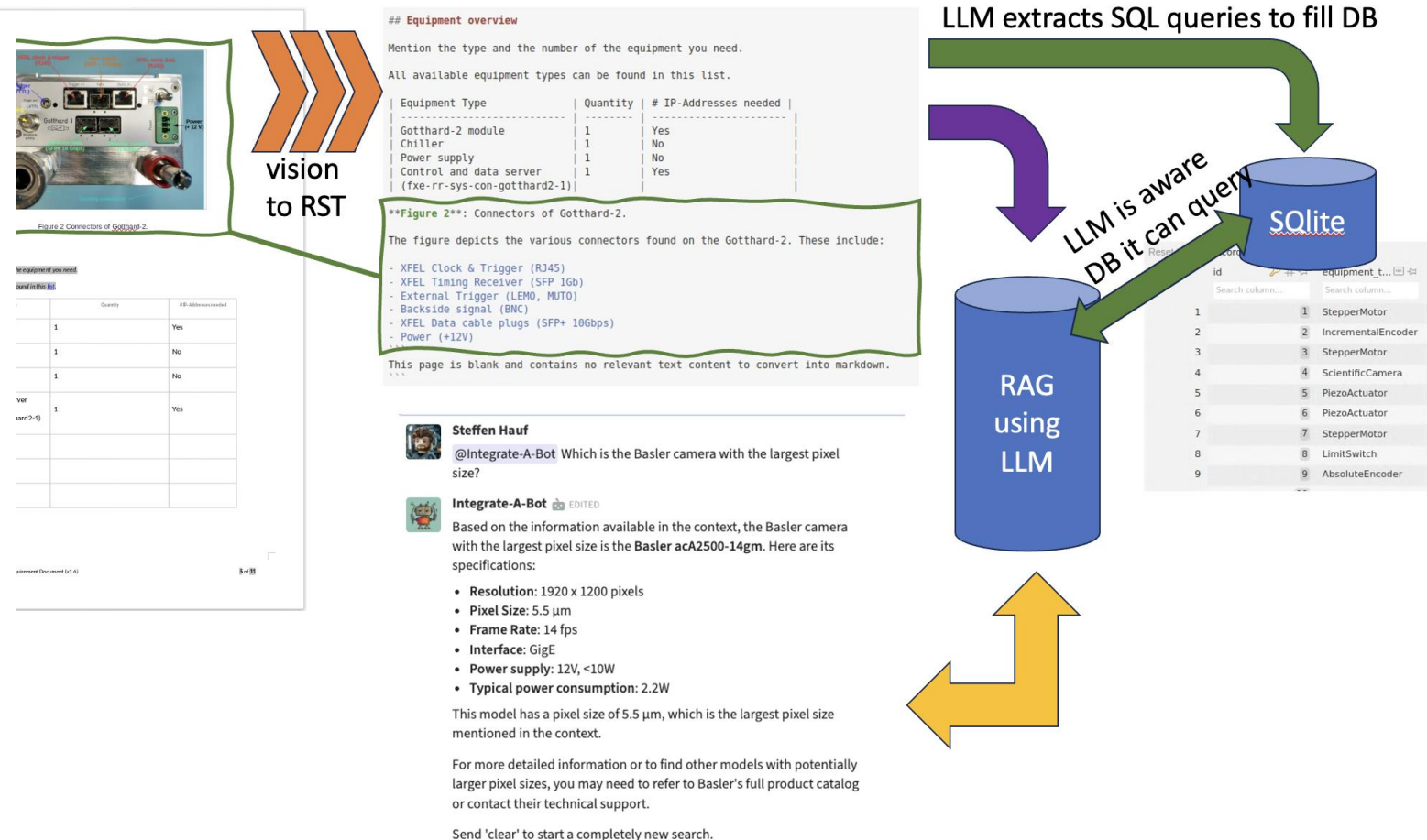
Contact: steffen.hauf@xfel.eu



Additional slides

Information Exploration

- A lot of information at EuXFEL is in form of PDFs and Word documents
- While practical, these are opaque information stores
 - Cannot efficiently search across many documents
 - Content is usually not very structured
- Vision-enabled LLMs are very efficient in converting this data into structured data that is **F**indable, **A**ccessible, **I**nteroperable, **R**eusable



Information Exploration

- A lot of information at EuXFEL is in form of PDFs and Word documents
- While practical, these are opaque information stores
 - Cannot efficiently search across many documents
 - Content is usually not very structured
- Vision-enabled LLMs are very efficient in converting this data into structured data that is **F**indable, **A**ccessible, **I**nteroperable, **R**eusable

Early prototype bot to enquire hardware available in the control system



Steffen Hauf

@Integrate-A-Bot 🤖 list all supported high-voltage sources with their voltage range in a tabular overview. Order them from highest supported voltage to lowest.



Integrate-A-Bot 🤖 EDITED

Below is a markdown table summarizing the high-voltage source equipment from the current database query. The entries are ordered from the highest known voltage rating to the lowest (or “not specified” where no voltage range is given):

Vendor	Model	Voltage Range
Stanford Research Systems	PS300m	Up to 30 kV
Glassman Europe	PS/EJ60P1N30	Not specified
Glassman Europe	PS/EJ010P120-28	Not specified
highvolt	HM345 series	Not specified
Iseg	HPn 20 157 QR52852	Approximately 2 kV (output is specified as
SEIFERT	ID 3003 - ISO-DEFLEX 3003	230 V AC (line supply; included in the quei

Notes:

- For the Iseg device, the specification “-2 kV” is given; here the absolute value is about 2 kV with negative polarity.
- For several devices (the two Glassman models and the highvolt HM345 series) no voltage range was provided in the database.
- The SEIFERT unit is included since it appeared in the query output, but its “230 VAC” rating indicates a mains-level supply rather than a typical high-voltage detector bias supply.

Suggested Policy: Lean & Draw From What's Working (at UniHH)

- There's rapid evolution going on in the field of LLMs
 - be lean and don't over-specify
- The University of Hamburg is providing LLM access to its students and staff based on a lean policy worked out together with the Landesdatenschutzbeauftragte of Hamburg
 - no need to reinvent the wheel for general access
 - fits onto 2(!) pages
- Foresee applications that will require direct API access
- Functional accounts anonymize input data
- Users may not enter and are trained to not enter sensitive and personal data. A classification system will help in this.

- Beschäftigte der ChatGPT-API, nur
- Die Nutzung and Ausnahmefällen aufzunehmen.

Die im folgenden genann

- Die Nutzung von Nutzung sollte m Risiken und ethis
- Die von einer KI g übernommen. Ein wissenschaftlich
- Zu Nachweiszwe Bildschirmausgal
- Die Regeln guter
- Die Nutzung von dargestellt (Zweck Zitationsregeln k
- Bei der Nutzung f Betracht gezogen
- Bestehende gese
 - Es werde eine KI üb Personen
 - Es werde Eingabe a
 - Die von e geschützt geschützt Verstoße:
- Wenn möglich ur durch entspreche ist bei UHH-GPT r
- Es werden keine l
- Werden eigene K externer Training

Die UHH bietet mit „UHH Chat-GPT-API. Im Rahmen

Univ.-Prof. Dr. Hauke He Hamburg, den 04.04.202

Dienstanweisung zur Nutzung des UHH-GPT-Angebots vom 04.04.2024

Der Einsatz künstlicher Intelligenz (KI) an der Universität Hamburg (UHH) in der Forschung, der Lehre und der Verwaltung ist mit Chancen und Risiken verbunden. Die UHH spricht sich für eine informierte und reflektierte Nutzung von KI aus und erwartet dabei die Einhaltung der Regelungen der guten wissenschaftlichen Praxis (GWP) und der geltenden rechtlichen Vorgaben insb. zum Datenschutz (DSGVO).

Die Nutzung von KI wird durch die disruptiven Entwicklungen und Angebote im Internet vereinfacht und für jeden Internetnutzer verfügbar. Geprägt wird die Entwicklung durch große Technologiefirmen und die Integration von KI-Systemen in immer mehr alltägliche Anwendungen. Zunehmend wird über die Authentizität von Texten, Bildern und anderen Medien und die Auswirkungen von KI auf die Gesellschaft diskutiert. Als Wissensorganisation müssen sich Universitäten in ihren Kernbereichen Forschung und Lehre (s. a. Empfehlungen zum Einsatz von KI in der Forschung und der Lehre) deshalb mit dieser Entwicklung auseinandersetzen und Position beziehen.

KI kann dabei Gegenstand der Forschung und der Lehre sein und/oder aber auch als Werkzeug für die Forschung und in der Lehre und Verwaltung genutzt werden.

Definition

Als KI bezeichnet man die Nachbildung menschlicher Intelligenz und/oder kognitiver Fähigkeiten mittels Computer-Systemen. Dabei kommen verschiedene Techniken zur Anwendung, wie z. B. maschinelles Lernen oder neuronale Netze.

Für die Nutzung an der UHH ist zu unterscheiden zwischen

- a. der Nutzung von **selbst betriebenen KI-Systemen**, z. B. zur Mustererkennung oder Datenanalyse oder auch als Gegenstand der Forschung, die mit vorzugsweise eigenen Trainingsdaten auf IT-Systemen der UHH betrieben werden und
- b. der **Nutzung von im Internet angebotene KI-Systemen**, die für verschiedene spezialisierte Aufgaben oder als universelles Werkzeug verwendet werden können.

Die unter a) genannten KI-Systeme sind in der Regel unkritisch, da sie in einem geschlossenen System betrieben werden. Allerdings können bei der Verwendung externer Trainingsdaten oder vortrainierter Modelle ebenfalls rechtliche Probleme auftreten, wenn in den Trainingsdaten urheberrechtlich geschützte Daten enthalten sind.

Die unter b) genannten KI-Systeme, gemeinhin auch als generative KI bezeichnet, da sie mit Hilfe von großen Modellen (z. B. Large Language Models (LLM)), maschinellem Lernen und sehr umfangreichen Trainingsdaten neue Inhalte generieren (erzeugen), erfordern bei ihrer Nutzung besondere Aufmerksamkeit.

Diese Dienstanweisung gilt deshalb insbesondere für die Nutzung von KI Angeboten im Internet und enthält Vorgaben für die Nutzung des von der UHH angebotenen Dienstes UHH-GPT, die von allen Beschäftigten der UHH zwingend einzuhalten sind.