CCIN2P3

cnrs

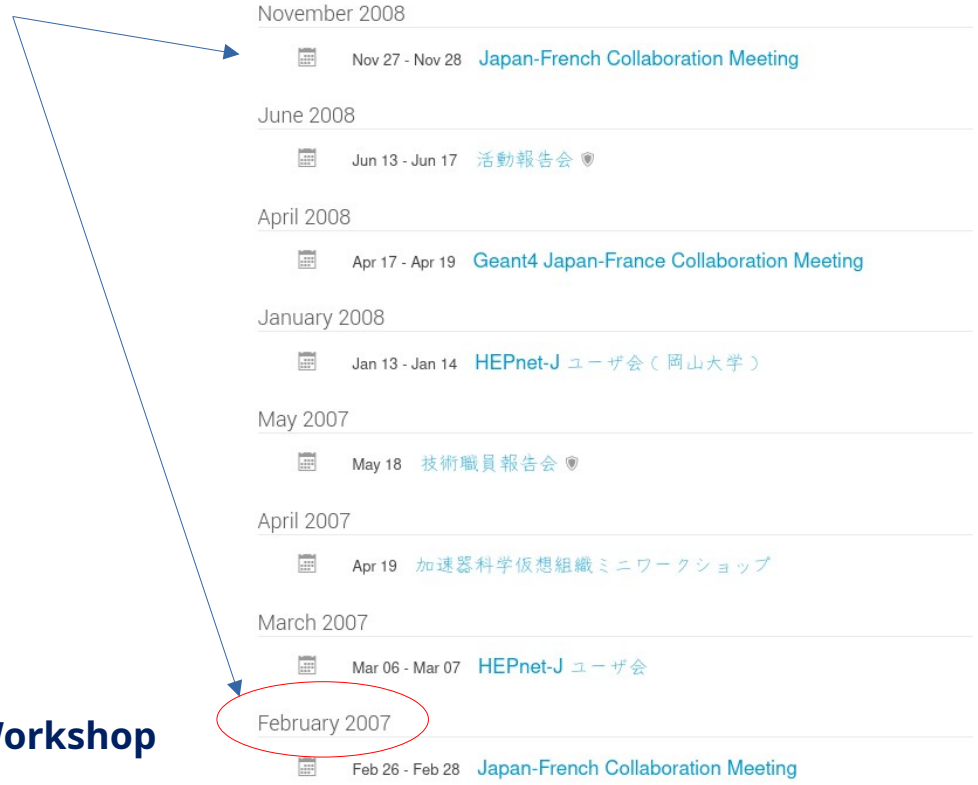
Centre de Calcul
de l'Institut National de Physique Nucléaire
et de Physique des Particules

Report on the COMP_05

R&D for the exascale computing environment

Introduction

- **Current members (*leader)**
  - **KEK CRC: Tomoe Kishimoto*, T. Nakamura, G. Iwai, K. Omori, J. Ueta, M. Saito**
  - **CC-IN2P3: Sébastien Gadrat*, F. Hernandez, S. Voisin, G. Marchetti, B. Rigaud, B. Chambon**
- **Collaboration between Japan and France started almost 20 years ago !**
- **Close collaboration between Japan and France is essential to address the new challenges arising from the evolution of computing for the particle and astroparticle experiments**



KEK CRC and CC-IN2P3

- **KEK CRC and CC-IN2P3 support a wide rande of experiments**
  - Several are common to both sites : BelleII, Comet, LHC, t2k, HyperK,…
- **Many <u>common topics of interests</u> and <u>challenges to address</u>**
- **In COMP_05, we chose to focus on**
  - Generative AI (LLM)
  - Identity federation
- **I will report on the main activities and progress since the last meeting**

Yearly F2F meetings

- **F2F meetings are unique opportunity to discuss on-going work at both sites.**
  - Meeting in Feb. at CC-IN2P3
  - Meeting in May at KEK

FJPPN — Japan-France workshop on computing technologies

Feb 10, 2026, 9:00 AM → Feb 11, 2026, 6:00 PM Europe/Paris

202 (CC-IN2P3)

Sébastien Gadrat (CC-IN2P3)

Description The goal of this workshop is to explore relevant technologies, exchange experience and share ideas among experts of both Japan and France organisations in several scientific domains.

This is the 7th edition of this workshop, which is organized annually in the framework and with the sponsorship of the France-Japan Particle Physics Laboratory. The agendas of previous editions are available:

- 2025
- 2024
- 2023
- 2019
- 2018
- 2017
- 2016
- 2015

FJPPN — Japan-France Workshop on Computing Technologies at KEK

May 14 – 15, 2026
KEK Computer North Bldg.
Asia/Tokyo timezone

Enter your search term

Overview
Timetable
Contribution List
Registration

The goal of this workshop is to explore relevant technologies, exchange experience and share ideas among experts of both Japan and France organisations in several scientific domains. This workshop is a satellite event of the annual workshop held as part of the FJPPN project. The agendas of previous editions are available:

- 2026
- 2025
- 2024
- 2023
- 2019
- 2018
- 2017
- 2016
- 2015

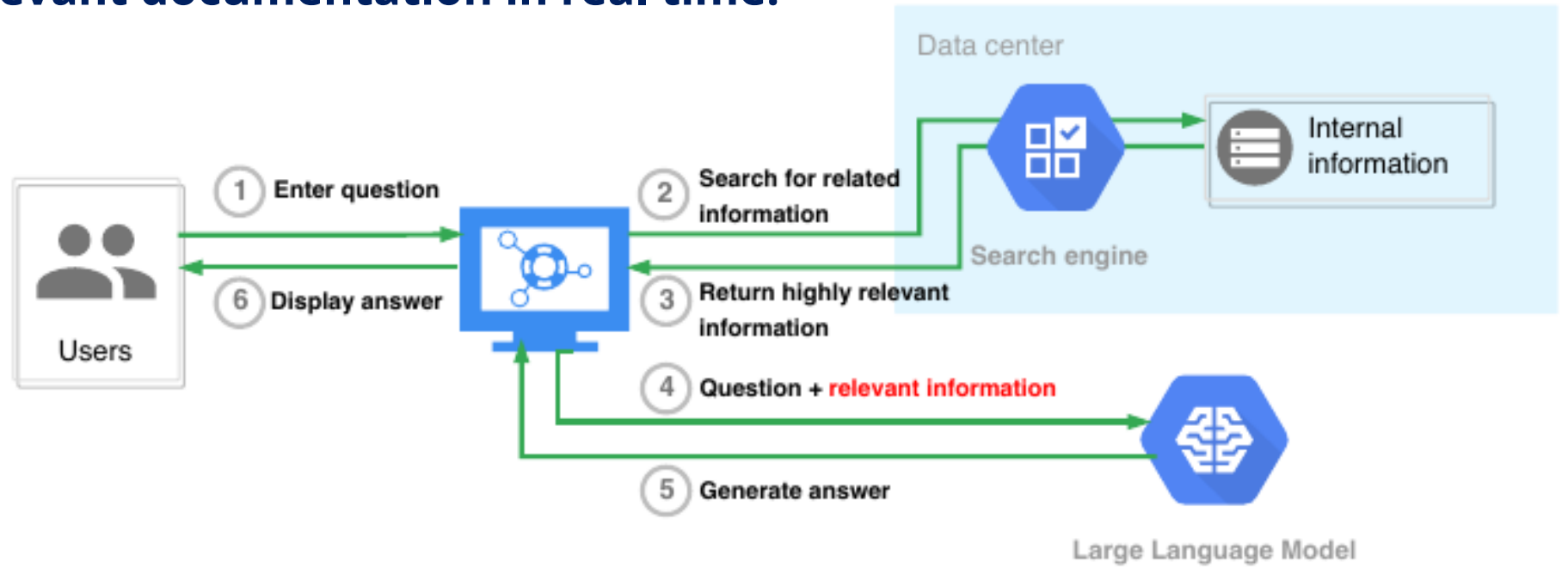
2 F2F meetings and 1 week hackathon!

Chatbots use in data center operations

User support associated workload reduction

Chatbots use in data center operations

- Many support requests involve information that already exists in the documentation but remains difficult for users to locate or interpret
- A LLM can help users, but does not have access to internal, nor up-to-date, information
- A RAG-LLM system, on the other side, can retrieve, aggregate, and contextualize the relevant documentation in real time.



Test system deployed (Proof-of-Concept)

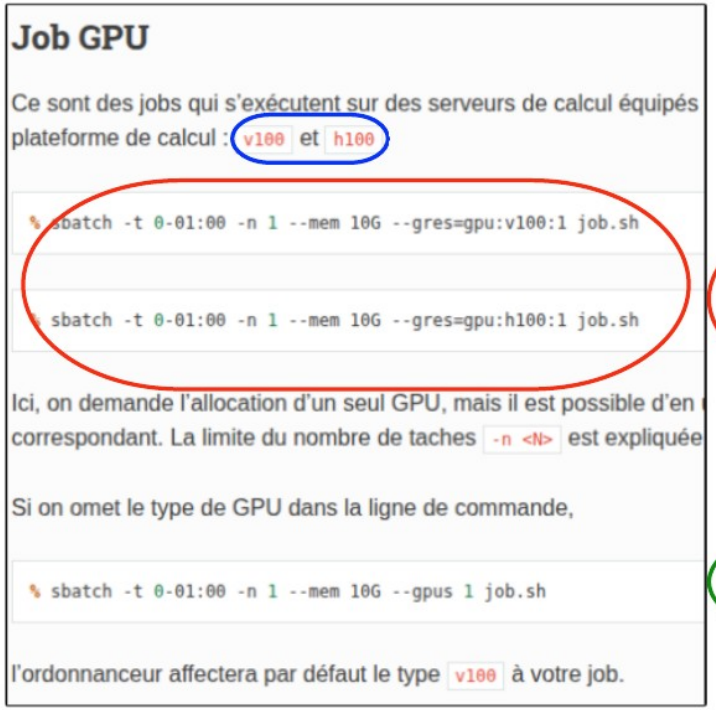
- **At CC-IN2P3 we set up a test system to validate the proof-of-concept of such chatbot**
  - running on premises Nvidia L40s GPU, and accessed through Jupyter notebooks (developed in Python)
  - based on the official documentation
  - is able to anwser either in french or english according to the langage used by the question
  - tested with various open source LLM models (based on Ollama CLI)

- **The PoC was successful, but we still need to improve a few things**
  - no friendly UI
  - several medium-sized LLMs were evaluated; further testing is needed to determine the optimal performance–energy efficiency trade-off
  - Automate RAG database workflow update

Quick demonstration

Démonstration

Comment puis-je spécifier le type de GPU que je veux utiliser ?



Job GPU

Ce sont des jobs qui s'exécutent sur des serveurs de calcul équipés
plateforme de calcul : `v100` et `h100`

```
% sbatch -t 0-01:00 -n 1 --mem 10G --gres=gpu:v100:1 job.sh
```

```
% sbatch -t 0-01:00 -n 1 --mem 10G --gres=gpu:h100:1 job.sh
```

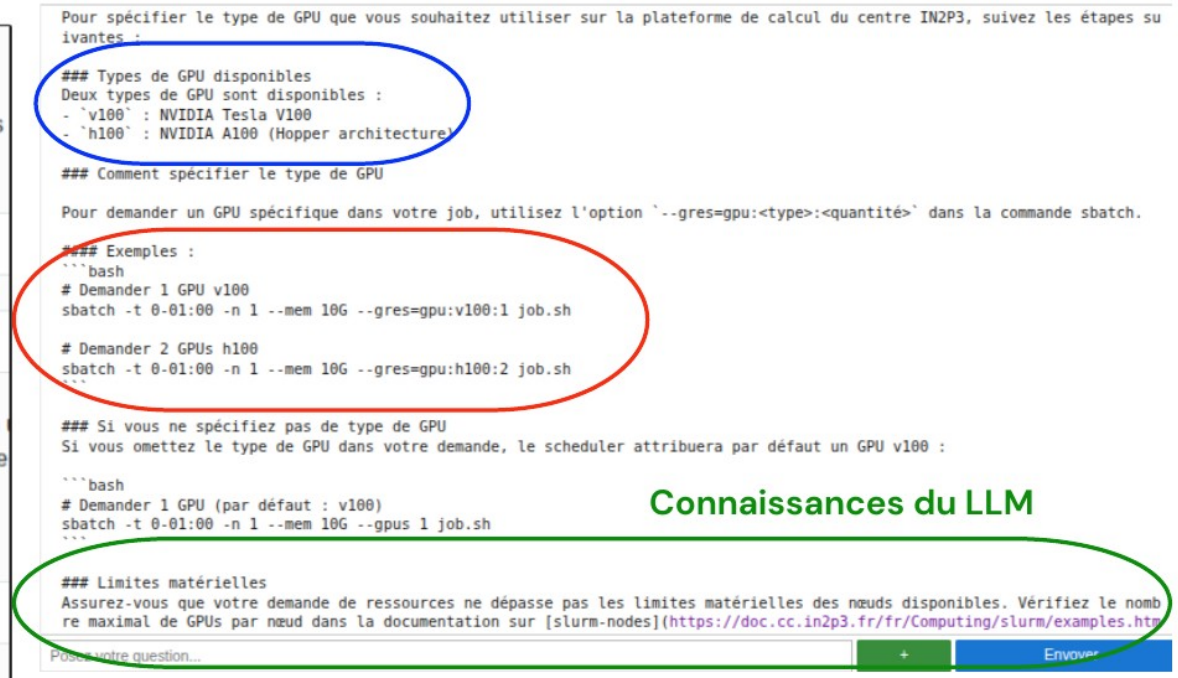
Ici, on demande l'allocation d'un seul GPU, mais il est possible d'en
correspondant. La limite du nombre de taches `-n <N>` est expliquée

Si on omet le type de GPU dans la ligne de commande,

```
% sbatch -t 0-01:00 -n 1 --mem 10G --gpus 1 job.sh
```

l'ordonnanceur affectera par défaut le type `v100` à votre job.

Documentation du CC-IN2P3



```
Pour spécifier le type de GPU que vous souhaitez utiliser sur la plateforme de calcul du centre IN2P3, suivez les étapes su
ivantes :

### Types de GPU disponibles
Deux types de GPU sont disponibles :
- `v100` : NVIDIA Tesla V100
- `h100` : NVIDIA A100 (Hopper architecture)

### Comment spécifier le type de GPU

Pour demander un GPU spécifique dans votre job, utilisez l'option `--gres=gpu:<type>:<quantité>` dans la commande sbatch.

#### Exemples :
```bash
# Demander 1 GPU v100
sbatch -t 0-01:00 -n 1 --mem 10G --gres=gpu:v100:1 job.sh

# Demander 2 GPUs h100
sbatch -t 0-01:00 -n 1 --mem 10G --gres=gpu:h100:2 job.sh
```

Si vous ne spécifiez pas de type de GPU
Si vous omettez le type de GPU dans votre demande, le scheduler attribuera par défaut un GPU v100 :

```bash
# Demander 1 GPU (par défaut : v100)
sbatch -t 0-01:00 -n 1 --mem 10G --gpus 1 job.sh
```

Limites matérielles
Assurez-vous que votre demande de ressources ne dépasse pas les limites matérielles des nœuds disponibles. Vérifiez le nomb
re maximal de GPUs par nœud dans la documentation sur [slurm-nodes](https://doc.cc.in2p3.fr/fr/Computing/slurm/examples.htm

Posez votre question...   +   Envoyer
```

Connaissances du LLM

Rapport de stage - BELZ Matteo- 9

The RAG system retrieves and aggregates relevant documentation while leveraging LLM knowledge to provide contextualized assistance.

AI hackathon @KEK

From Tomoe's report last year at Nantes joint workshop…

- To strengthen the partnership, we request funding for:
  - **IN2P3-CC → KEK-CRC visit in Autumn** to discuss and jointly work on AI-related activities (AI hackathon)

A new initiative

- **Thanks to a KEK grant, Sybille was able to come to KEK for one week in May (last week)**
- **We had 1-week hackathon to implement a RAG-LLM system on a Nvidia DGX Spark (providing a Grace Blackwell GPU)**

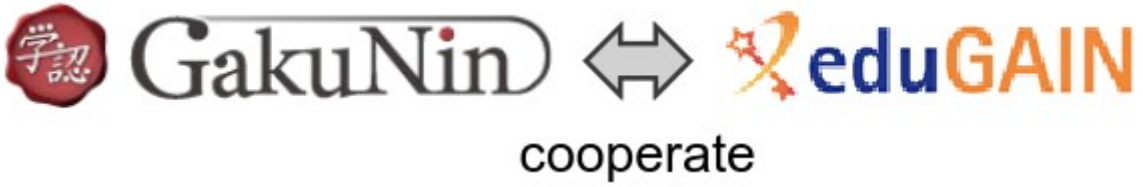
- **1-week was a bit short, but Sybille and Tomoe were able to validate a RAG PoC (KEK CRC internal documentation) on the DGX system**
- **Further tests are however needed to better deal with the limitations of the system (especially memory, as medium LLM systems require high memory to run)**

Identity federation

Enabling seamless cross-service authentication using institutional credentials

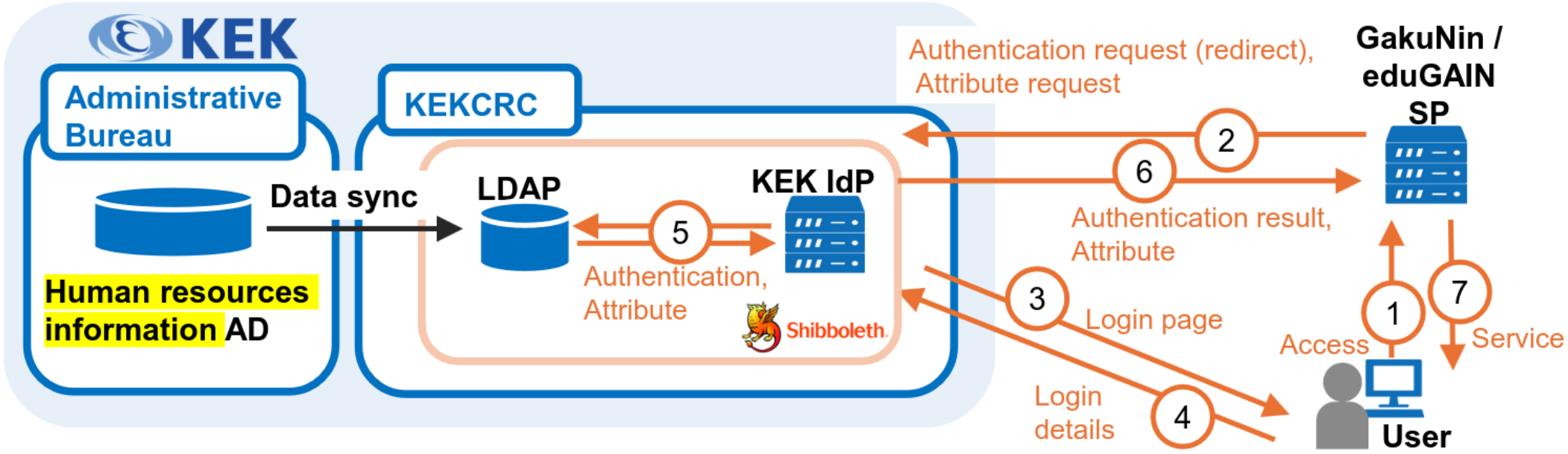
Identity federation

- Identity Federation provides seamless access to distributed computing services across collaborating institutes
- KEK CRC and CC-IN2P3 participate in several international particle and astroparticle physics collaborations relying on shared computing and data infrastructures, as well as collaborative tools
- Identity federation helps to achieve this, by allowing users to authenticate using their home institution credentials and seamlessly access to all required services
- Interoperability between Identity Providers (IdP) and Service Providers (SP) is therefore a key requirement for cross-site scientific computing collaborations
- Academic Identity federation



Build of KEK's Identity Provider

- KEK built a Identity Provider (Shibboleth based) in order to join GakuNin
- This allows KEK to join GakuNin (and therefore all SP accessible from it)
- GakuNin being able to communicate with eduGAIN, this enables KEK users to smoothly access CC-IN2P3 collaborative tools (such as GitLab)

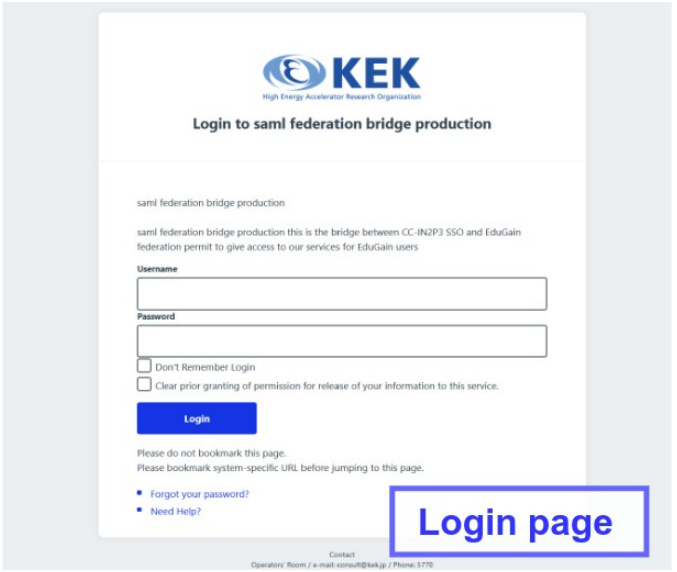


Next steps

- First step achieved: set up an Identity Provider and joined GakuNin

- But still significant work remain to reach the final objective

- Need to work on the SSO (Single Sign On) to allow seamless access to all services provided by KEK CRC
  - Currently technology, i.e. Shibboleth, might not be enough
  - At CC-IN2P3, we are using Keycloak



- Need to make it work the other way around (CC-IN2P3 → KEK CRC)
  - Service Provider may require different information from the Identity

Summary & Outlook

- **Since the last meeting, we managed to have 2 F2F meetings, and 1-week hackaton at KEK**
  - In-person meetings (especially the hackaton) foster collaboration and faster progress
- **We achieved some major milestones in both projects, almost significant work remain ahead**
- **On Generative AI (LLM)**
  - Further testing is required to evalute the trade-off performance vs resources
  - Additional content should be integrated such as optimized code and practical examples to better assist users in developing efficient application
- **The AI landscape is evolving rapidly, we need to follow up**
  - Current trend is AI agents which run in background, and are capable of performing various tasks (this however arises security and trust concerns)
- **On identity federation**
  - Develop a KEK SSO
  - Ensure seamless access for french users to KEK CRC services through federation

Outlooks & Plans for 2027

- The combination of focused technical work sessions (hackaton) and in-person discussions (workshop) proved highly effective for advancing on our projects

- We therefore would like to repeat this initiative next year for both teams
  - We plan to have 2 workshops, at CC-IN2P3 and at KEK
  - We would like to be able to stay ~1 week before or after the workshop
  - In the 2026 application, we therefore asked for 15 days funding for 2/3 persons
  - We will combine KEK workshop, technical work sessions at KEK and Joint TYL/FJPPN and FKPPN workshop (Korea) in the travel to Japan

ごせいちょうありがとうございました。ごしつもんはありますか?

Thank you ! Questions ?