

New proposal: R&D for the exa-scale computing environment

Team:

France: [Sébastien Gadrat](#), F.Hernandez, G.Marchetti, S.Voisin, G.Mainetti, I.Magroune

Japan: [Tomoe Kishimoto](#), T.Nakamura, G.Iwai, K.Omori, M.Saito

Introduction

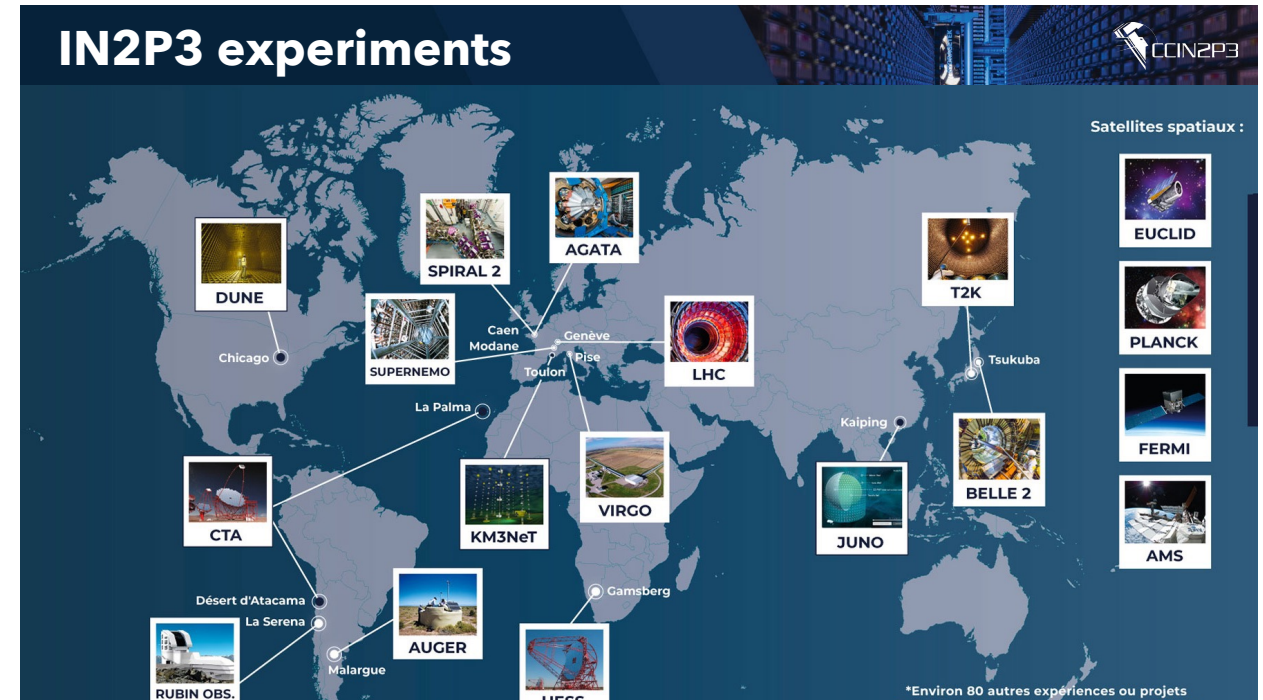
- This new proposal follows COMP_04 (and COMP-03) as its successor
- COMP_04:
 - “Evolution of the computing environment for high-energy and astro-particle experiments”
 - The project was led by [Fabio Hernandez](#), [Renaud Vernet](#), [Sebastien Gadrat](#), and [Tomoaki Nakamura](#) from FY2019 to FY2024
 - The close collaboration between IN2P3 Computing Center (CC-IN2P3) and KEK Computing Research Center (KEK CRC) in computing

→ Let me report the achievements of COMP_04 in FY2024



CC-IN2P3 and KEK CRC

- CC-IN2P3 and KEK CRC support a wide range of experimental projects
 - Including BelleII, LHC, T2K, etc
- Many common challenges to overcome
 - Both hardware and software areas
- Shared interests regarding new technologies
 - Machine learning(ML), ID federation, etc



→ Exchange of experience, and know-how is quite essential !!

Yearly F2F workshop

- F2F workshops are unique opportunity to discuss on-going work and new technology at both centers
- Topics of 2025 workshop@Lyon were:
 - Status report from CC-IN2P3 and KEK CRC
 - News from LSST experiment
 - Open access project at KEK
 - AI session ← Focus of this presentation

<https://indico.in2p3.fr/event/35206/>

FJPPL — Japan-France workshop on computing technologies

Feb 18, 2025, 9:00 AM → Feb 19, 2025, 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.

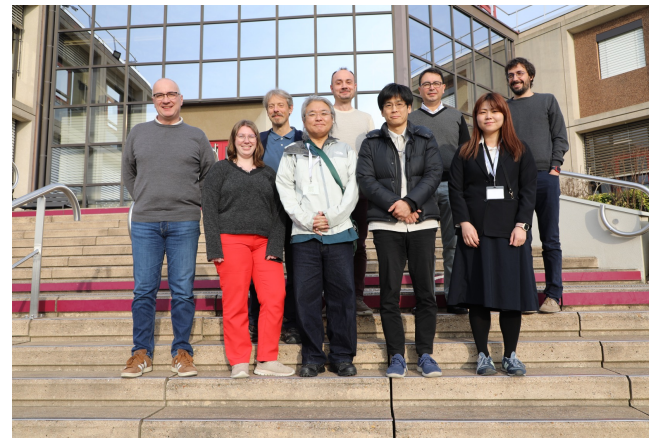
This is the 7th edition of this workshop, which is organized annually in the framework and with the sponsorship of the [France-Japan Particle Physics Laboratory](#). The agendas of previous editions are available:

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

COMP_04

COMP_03

Workshops since 2015!



加速器だから見える世界。



AI session

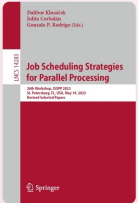
Home > Job Scheduling Strategies for Parallel Processing > Conference paper

An Efficient Approach Based on Graph Neural Networks for Predicting Wait Time in Job Schedulers

Conference paper | First Online: 15 September 2023
pp 137–154 | [Cite this conference paper](#)

Access provided by High Energy Accelerator Research Organization

[Download book PDF](#) [Download book EPUB](#)



Job Scheduling Strategies for Parallel Processing
(JSSPP 2023)

➤ ML has been a focus of joint collaborative efforts between CC-IN2P3 and KEK CRC

➤ A successful application for estimating job waiting time in batch schedulers

➔ **Published in Lecture Notes in Computer Science**

WED, FEBRUARY 19

8:30 AM → 9:00 AM Welcome coffee 30m

9:00 AM → 10:00 AM AI session 1h

9:00 AM Large Language Model
Speaker: Mr Mehdi Hennequin
LLM.pdf

10:00 AM → 10:30 AM Coffee break 30m

10:30 AM → 12:00 PM AI session 40m

10:30 AM A study of foundation models for event classification in collider physics
Speaker: Tomoe Kishimoto (High Energy Accelerator Research Organization)
2025_02_17.pdf

11:10 AM MLLMs in Scientific Research and High-Performance Computing 50m
Speaker: Imed MAGROUNE
MMLMs_state_of_t...

AI session agenda

➤ New application has been investigated:

➤ The idea is to apply the concept of foundation models to collider physics

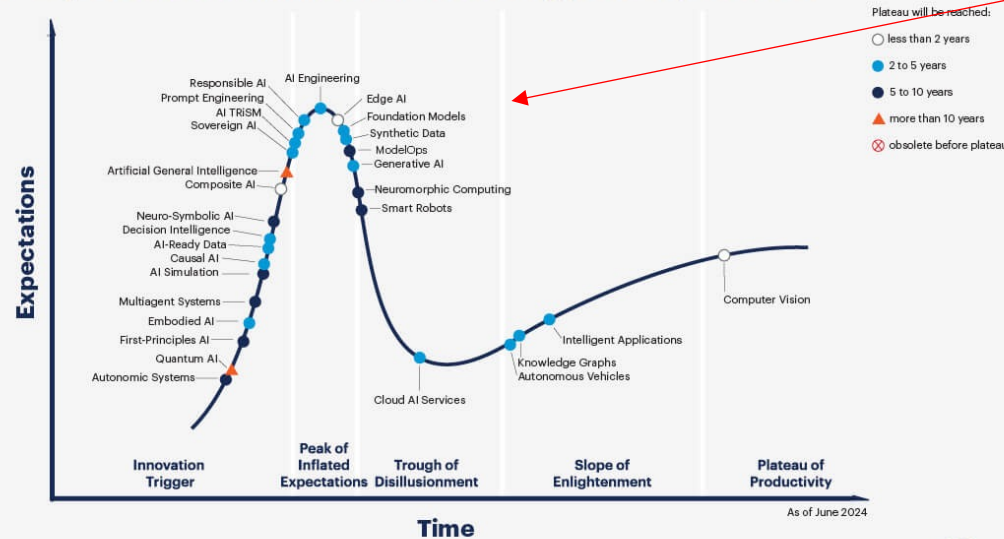
➤ A potential approach to reduce computing resource usage in future collider experiments

加速器だから見える世界。



Foundation models

Hype Cycle for Artificial Intelligence, 2024



Source: Gartner
Commercial reuse requires approval from Gartner and must comply with the
Gartner Content Compliance Policy on gartner.com.
© 2024 Gartner, Inc. and/or its affiliates. All rights reserved. GTS_3282450

Gartner®

[Gartner.com](https://www.gartner.com)

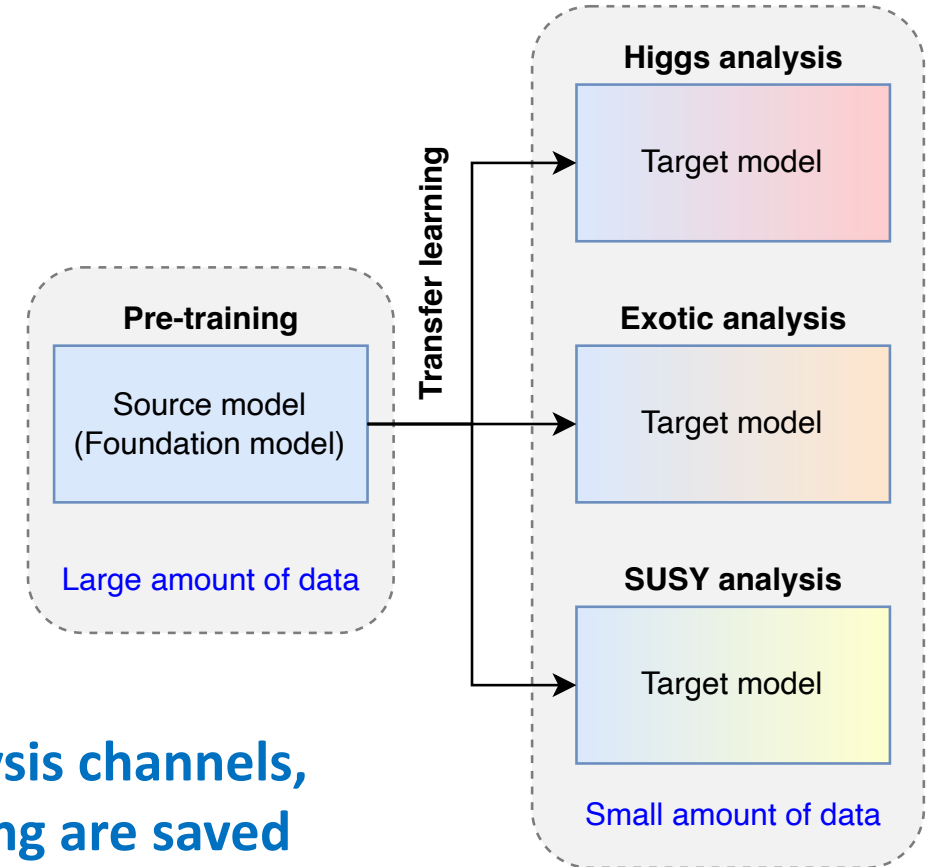
- “Foundation models” is one of the keywords for AI
- Pre-training using a larger amount of “unlabeled” data
- Fine-tuning for a target application (transfer learning)

→ Q: Is the concept of foundation models beneficial to collider physics?

Use case of physics analysis

- Many analysis channels in collider physics
 - Higgs, Exotic, SUSY, etc
 - Dedicated Deep learning models are trained from scratch for each analysis channel
 - Large amount of training data (MC) for each channel

→ If transfer learning can be applied to different analysis channels, computing resources for MC simulations and DL training are saved



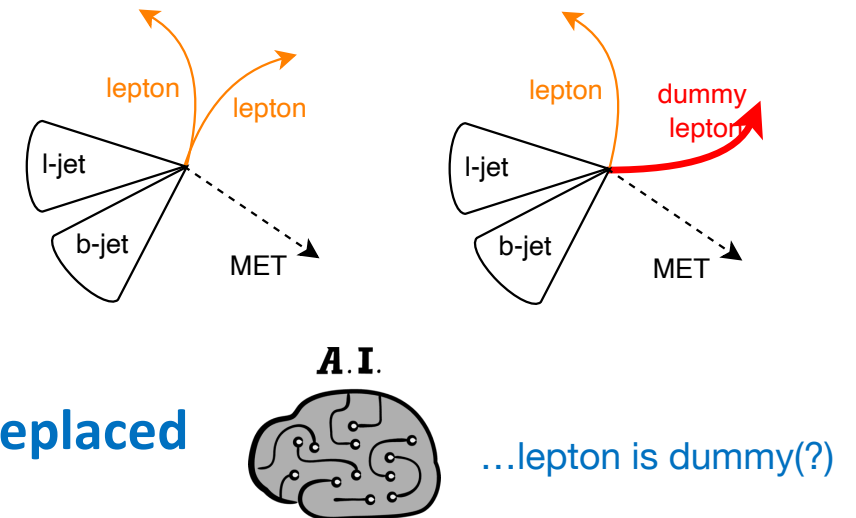
Pre-training strategy

- Only low-level features of each object (4-vector+charge) are used as inputs
- Real collision data (CMS open data) are used in pre-training
 - Self-supervised learning is employed to handle the unlabeled data

- Pre-training strategy

- An object is randomly replaced with a dummy object when preparing mini-batch

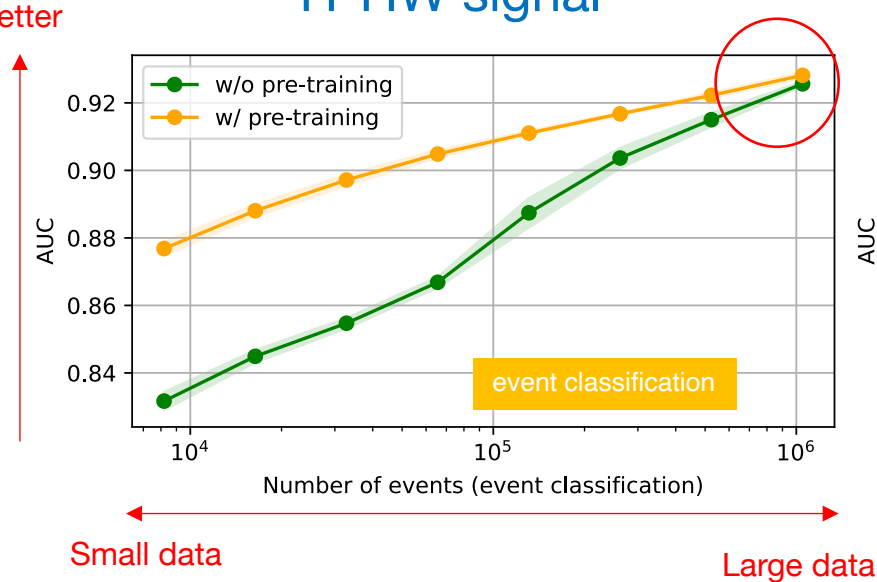
→ DL model is trained to predict what type of object was replaced



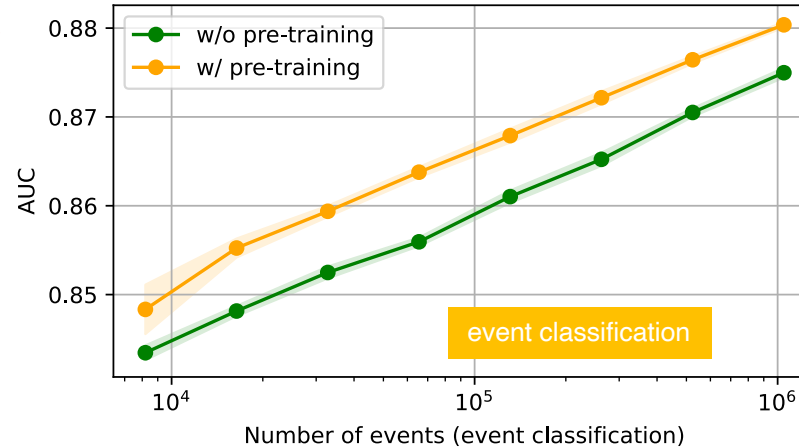
AUC of event classification

- The concept is examined using “Event classification” problem : signal vs background

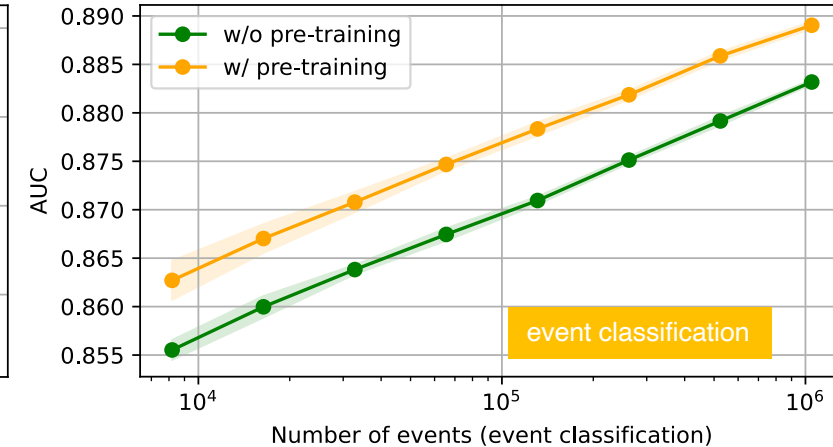
H⁺HW signal



ttH (1lep) signal



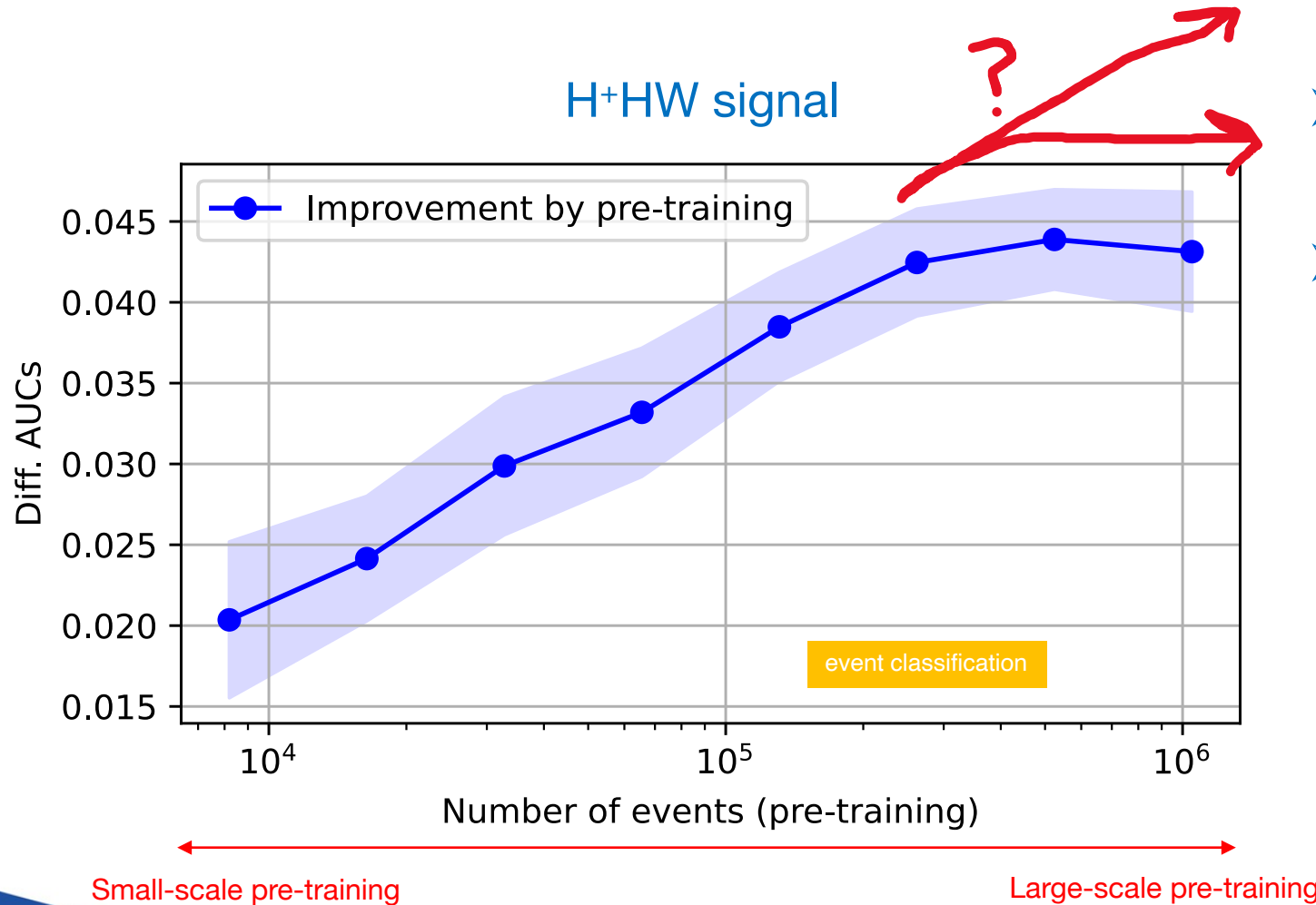
ttH (2lep) signal



- The improvements are confirmed for different signal events

➔ The pre-trained model (foundation model) generalizes well

Scaling raw



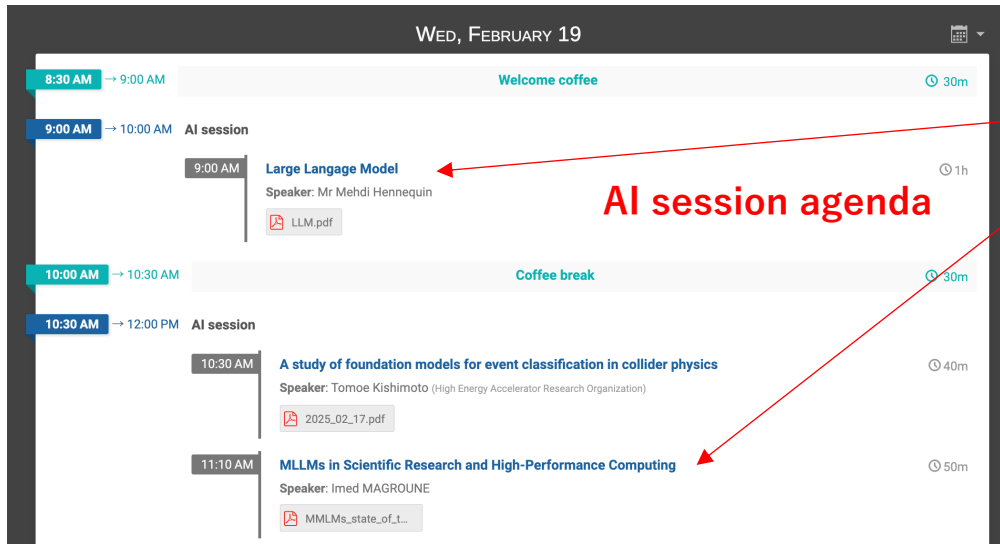
➤ The scaling behavior with respect to the number of events in pre-training

➤ **Future works:**

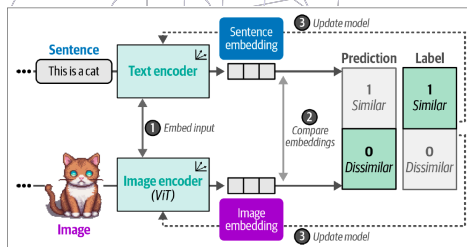
- Need to check with larger models and larger data
- Investigate new data augmentation approach
- Explainability of AI: what kind of knowledge is learned during pre-training and utilized in downstream tasks...?

Large Language Models

- We also discussed Large Language Models (LLM) and their applications in our fields during the workshop@Lyon
- User support for the efficient use of computing resources is a key challenge shared by CC-IN2P3 and KEK CRC
 - Distributed computing, HPC cluster, containerized workflow, Networks, etc



How Can CLIP Generate embeddings to drive the generation of images



From Foundation models to smaller LLMs



Jetson Orin Nano Developer Kit NVIDIA

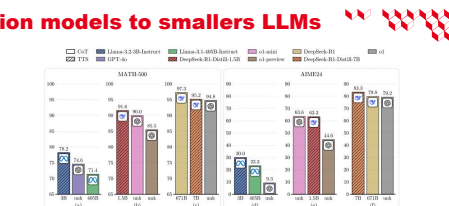


Figure 1: Comparison between the performance of smaller LLMs compute-optimal TTS and that of larger LLMs CoT on MATH-500 and AIME24. (a) & (d) Llama-3.2-3B-Instruct surpasses Llama-3.1-405B-Instruct and GPT-4o on MATH-500 and AIME24; (b) & (e) DeepSeek-R1-Distill-1.5B outperforms o1-preview on MATH-500 and AIME24, and surpasses o1-mini on MATH-500; (c) & (f) DeepSeek-R1-Distill-7B beats o1 on MATH-500 and AIME24, and exceeds DeepSeek-R1 on AIME24.

* Work done during an internship at Shanghai AI Laboratory
† Corresponding authors: Bingqi Qi (qibingqi@sjlab.org.cn), Bowen Zhou (zhoubowen@sjlab.org.cn)

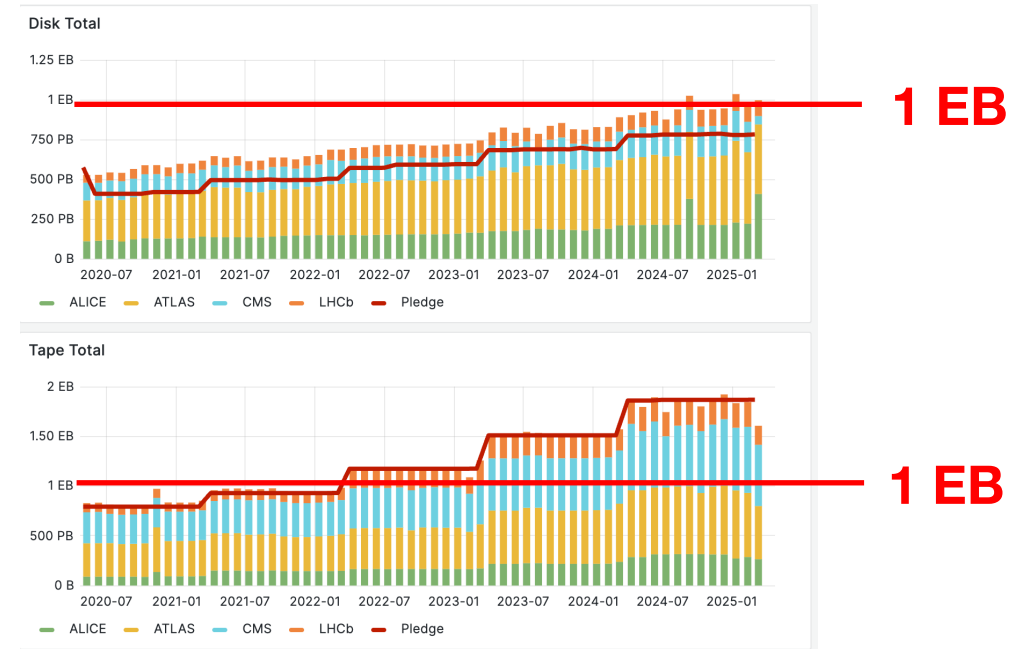
Can 1B LLM Surpass 405B LLM? Pioneering Compute-Optimal Test-Time Scaling. <https://arxiv.org/abs/2502.06703>

加速器だから見える世界。



New proposal

- Multiple approaches for introducing brand-new technologies are necessary for the evolution toward the exa-scale computing environment
- Refine and extend the knowledge gained from COMP_04
- 3 topics selected for this year:



- 1. Study of foundation models for event classification in collider physics**
→ This topic has been discussed in previous slides
- 2. Research and application of chatbots (based on LLMs) in High Energy Physics**
- 3. Research on enhancing the convenience of IT services based on identity federation**

Chatbots in HEP

- User support in data center operations requires access to **internal information**
 - System configuration, Past tickets, etc
- Normal chatbots can't answer such questions (of course!)

→ **Retrieval-Augmented Generation (RAG)** is a promising approach

ChatGPT 4o >

Could you please provide the IP address range for KEK's DC cluster?

The specific IP address range for KEK's Data Center (KEKCC) cluster is not publicly disclosed in the available documentation. However, access to KEKCC resources is managed through designated servers: WEB2.INFN.IT

- **Access Servers:** sshcc1.kek.jp and sshcc2.kek.jp
- **Work Server:** login.cc.kek.jp WEB2.INFN.IT KEKCC.KEK.JP

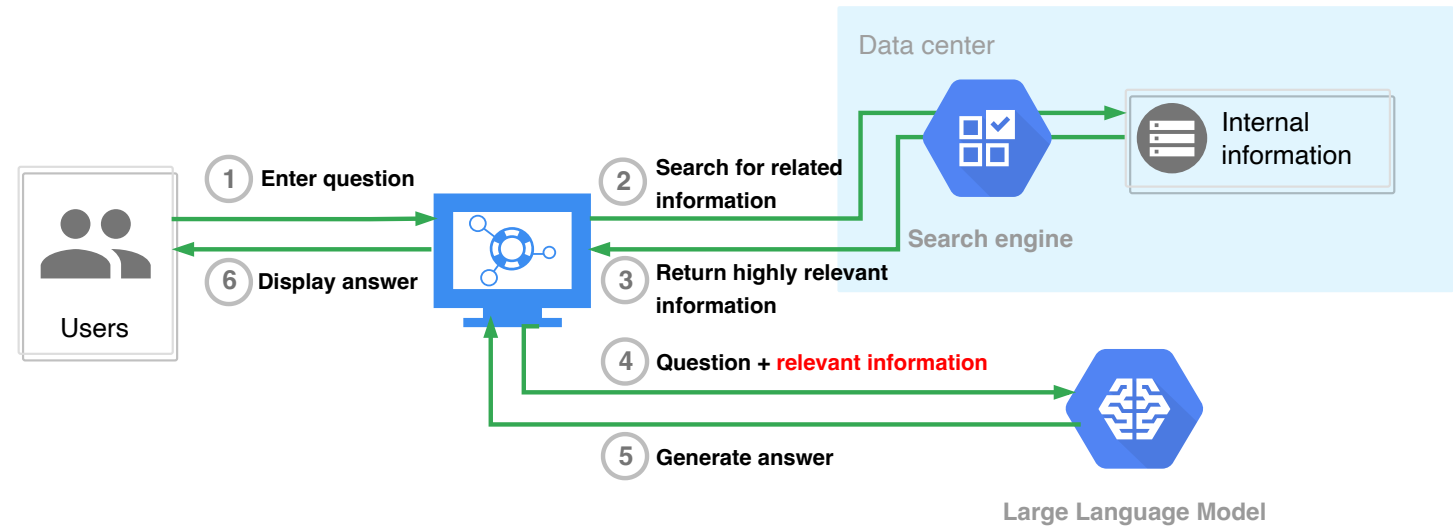
These servers facilitate secure connections for authorized users. Notably, direct access to the work server (login.cc.kek.jp) is restricted to clients with registered IP addresses. This implies that KEK maintains a list of approved IP addresses for security purposes. To gain access, users typically need to submit their machine's IP address and hostname for registration. This process is outlined in KEK's support documentation. INDICO.BELLE2.ORG KEKCC.KEK.JP WEB2.INFN.IT

If you require the exact IP address range for network configuration or firewall settings, it is recommended to contact the KEK Computing Center directly or consult with your KEK project liaison. They can provide the necessary information tailored to your access needs.

情報源

RAG

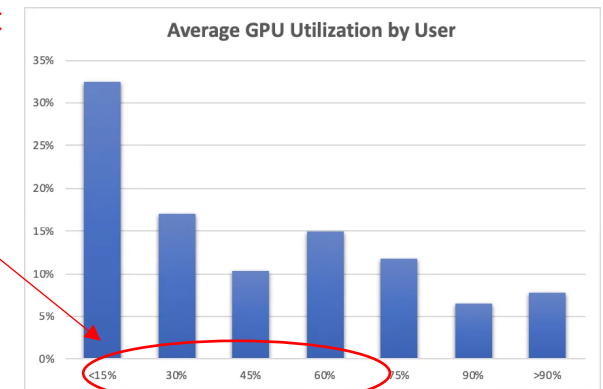
- Relevant information is retrieved and fed into LLMs for processing
 - Official documents, past ticket, e-mails, etc



→ Develop a chatbot for user support, including code generation optimized for our computing infrastructure

- Related research topics:
 - Methods for searching relevant information efficiently
 - Methods for fine-tuning or distillation to enable on-premise deployment of LLMs

Code optimization is necessary, but it heavily depends on the system



[Link to article](#)

ID federation

- ID federation plays a crucial role in the efficient deployment of IT services (including chatbot)
 - Users are able to use services provided by other institutions with the credentials from their home institution

→ Reduce the effort required for account management

- Academic ID federation:
 - European ID federation → eduGAIN
 - Japanese ID federation → GakuNin
 - GakuNin can federate with eduGAIN



→ Plan to deploy a test service for verifying authentication and the exchange of user attributes

Migration: X.509 to Token

- ID federation is particularly critical for “Grid computing”
- The Grid authentication infrastructure is migrating from X.509 certificates to token-based authentication
 - KEK CRC is deploying a token issuer service (INDIGO IAM) for BelleII experiment
 - But, a user registration procedure for INDIGO IAM is not established yet...

→ Managing user identities and accounts for all collaborators with the security level required by Grid computing^(*) is a significant challenge

→ We need ID federation!

(*) Identity Assurance Level2 must be referred to F2F meeting or human resource DB

Summary

- Information exchange is highly valuable for both centers
- In the new proposal, we plan to focus on the application of AI technologies and ID federation
 - Milestones for FY2025: Development of an initial chatbot prototype and deployment of a test service for ID federation
- 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)
 - **KEK-CRC → IN2P3-CC visit in Spring** to exchange information on computing technologies (annual workshop)

A new initiative

Backup

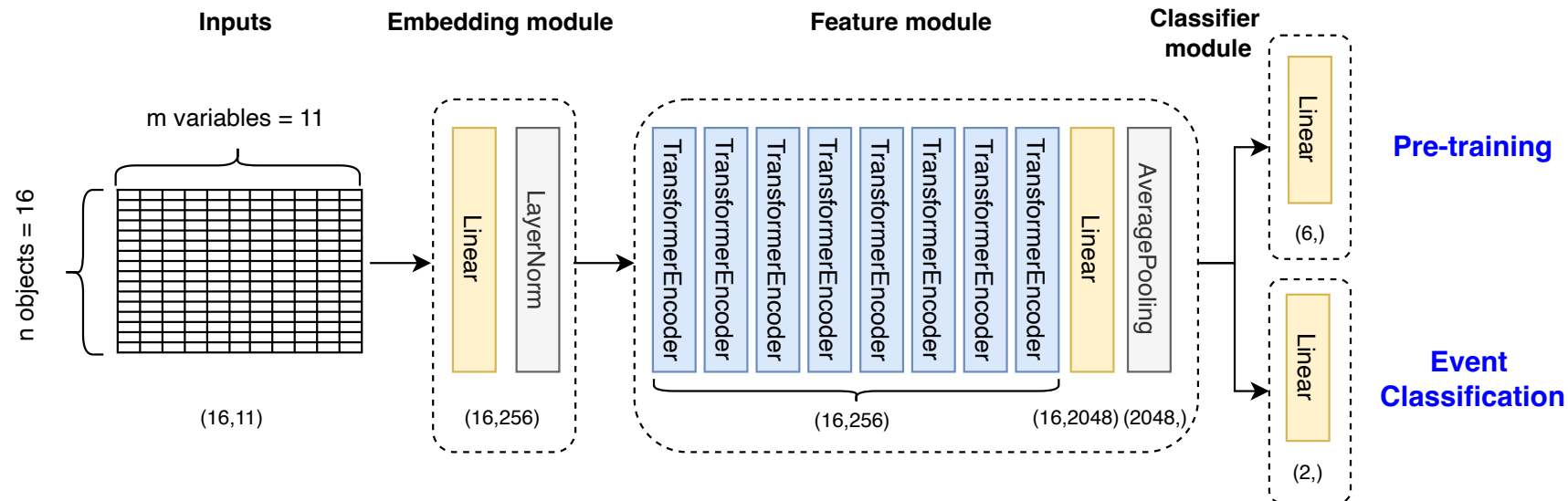
Datasets (CMS Opendata)

		Selections	# of events
Pre-training →	Collision data	lepton ≥ 1 + jets ≥ 2 + bjets ≥ 1	$\sim 10^6$
Event classification {	H+tb[ref.] vs ttbar+jets	lepton ≥ 1 + jets ≥ 4 + bjets ≥ 1	$\sim 10^6$
	H+HW[ref.] vs ttbar+jets	lepton ≥ 1 + tau ≥ 1 + jets ≥ 3 + bjets ≥ 1	$\sim 10^6$
	ttH[ref.] vs ttbar+jets	lepton ≥ 1 + jets ≥ 4 + bjets ≥ 2	$\sim 10^6$
	ttH[ref.] vs ttbar+jets	lepton ≥ 2 + jets ≥ 2 + bjets ≥ 1	$\sim 10^6$

- Pre-training is performed using collision data (unlabelled data) based on the foundation model concept
 - $\sim 10^7$ events are available after the selection, but only $\sim 10^6$ events are used
 - NVIDIA A100: $\sim 10^4$ events/sec (10^7 events / $10^4 \times 500$ epochs = 138 hours)

DL model

- Transformer encoder is employed:
 - ~11M trainable parameters



- Weight parameters of embedding and feature modules are transferred and fine-tuned
- Classifier module is always trained from scratch