



Обзор международного совещания Hepix-Spring-2025

<https://indico.cern.ch/event/1477299/>

Докладчик А.Е. Шевель

План

- Сведения о НЕРіХ
- Представленные крупные зарубежные лаборатории
- Компьютинг: состояние, тенденции, проблемы, ближайшее будущее
- Технологии машинного обучения и наша RAG архитектура, дальнейшие применения.

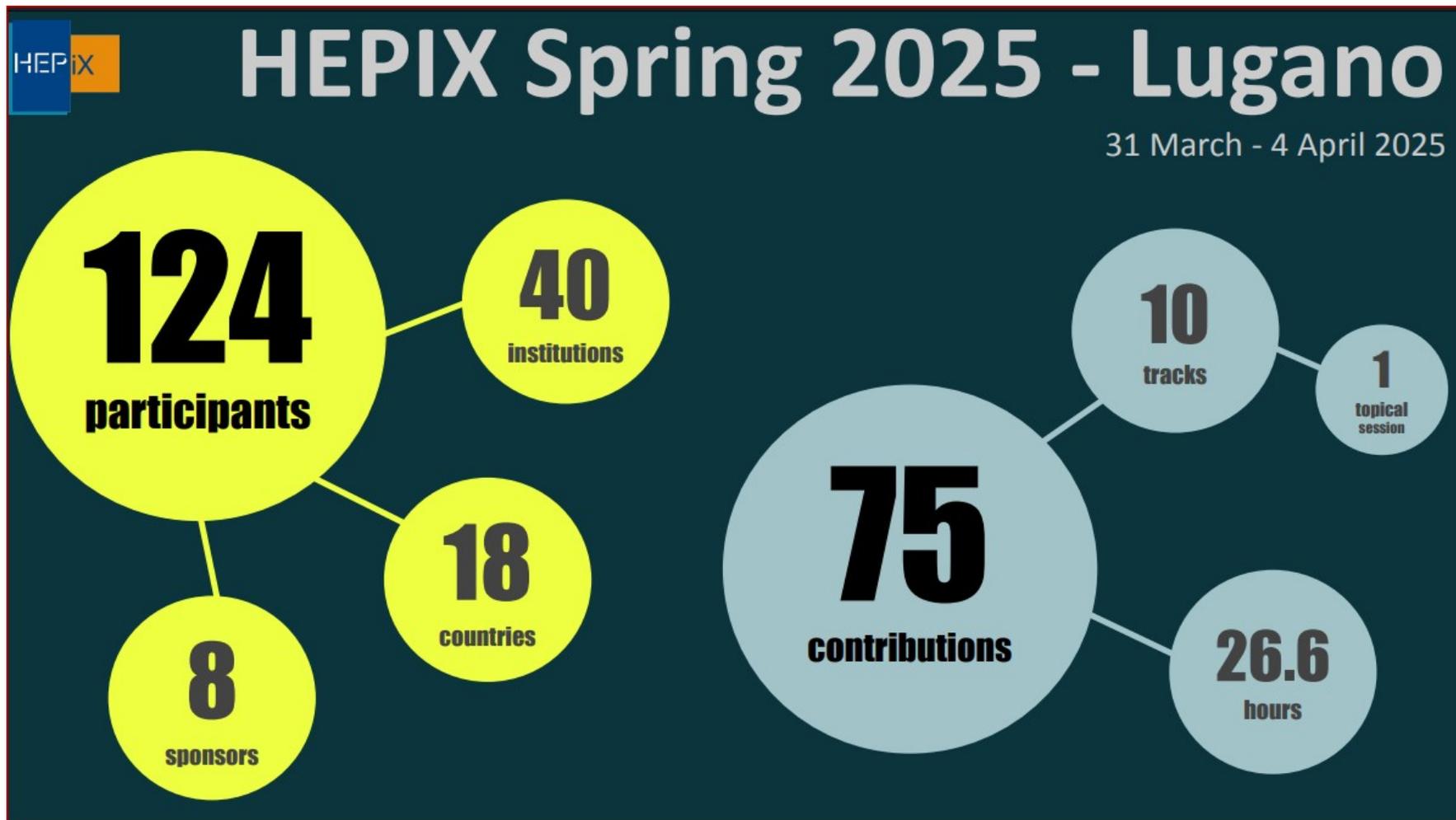
Что есть HEPiX

- The HEPiX forum brings together worldwide **Information Technology staff**, including system administrators, system engineers, and managers from the **High Energy Physics and Nuclear Physics laboratories and institutes**, *to foster a learning and sharing experience between sites facing scientific computing and data challenges*. Its semi-annual meetings are an excellent source of information and sharing for IT experts in scientific computing.

Научная программа Herix Spring 2025

- Новости крупных вычислительных установок
- Сеть и безопасность
- Хранение и управление данными
- Облачные технологии, виртуализация и оркестровка, операционные системы
- Вычислительные системы и пакетная обработка данных
- Экологическая устойчивость, непрерывность вычислений
- Программное обеспечение и услуги для эксплуатации вычислительных систем
- Средне-долгосрочная эволюция вычислительных систем
- Разное

Статистика Hepix Spring 2025



Организатор: Swiss National Supercomputing Centre



SNSC

- SNSC operates a Tier-2 grid site for the Large Hadron Collider (LHC)
- at CERN for three experiments: ATLAS, CMS and LHCb.
 - Over 10'000 cores and 10 PB of data at SNSC

Alps

- Alps is an HPE Cray EX supercomputer
- Multi-phase installation, started in 2020
- Some specs
 - 2688x GH200 nodes
 - 1024x MC nodes (AMD Rome 7742)
256/512GB RAM
 - 144x nVIDIA A100 GPU nodes
 - 128 AMD Mi300A CPU+GPU nodes
 - 24x AMD Mi250x GPU nodes
 - Slingshot network
 - Two available zones (HA, non-HA)
 - 100% liquid cooled (to the chips themselves)

Крупные лаборатории

- КЕК (штат ~1100) - LHC, Belle-2 и другие
- DESY (штат более 3000) - Cherenkov Telescope Array Observatory (CTAO), LHC, другие
- BNL (штат ~3000) — ATLAS, sPhenix+STAR, future Electron-Ion Collider (EIC) и другие
- IHEP (штат ~1000) — LHC + более 10 других экспериментов.
- CERN (45K accounts) — LHC
- RAL (штат ~1200) - LHC

Рост объёмов сохранённых данных

- CERN — достигло
 - 1000 PB (1 EB) in EOS (диски)
 - about 1 EB (ленты)
- BNL — достигло
 - ~300 PB (диски)
 - Ленты 345 PB

Модернизации аппаратных ресурсов

- BNL
 - 555 servers for sPHENIX
 - 75 servers for ATLAS Tier-1
- IHEP — внешняя сетевая
коннеktivность 100 Gbit, внутри IHEP
сетевая коннеktivность между
корпусами/установками достигла 800
Gbit.
- Модернизация систем распределения
электропитания во всех крупных ЦОДах

Модернизация мониторинга

- **BNL** - внедрение новой системы мониторинга компьютерной инфраструктуры с использованием
 - Prometheus, Grafana Mimir, Opensearch, Opensearch Dashboards
 - Testing Mattermost integration for alerting
 - Testing BNL ITD Gravwell for integration to reduce storage overhead and data duplication

Технические тенденции

- Сети
 - Переход всех лабораторий на IPv6
 - SkiTags.org (тегирование сетевых пакетов в передаче данных между лабораториями в международном трафике)
- Хранилища данных
 - CEPH — DESY, SNSC, RAL
 - Lustre — BNL, IHEP, SNSC
 - EOS - CERN, IHEP
 - ZFS — BNL
 - FAIR (Findable, Accessible, Interoperable, Reusable) — CERN (Zenodo), BNL (InvenioRDM)
- Machine Learning
 - CERN
 - DESY
 - IHEP
 - BNL
- Контейнеризация ПО (Rancher, Helm, Kuberflow, Harvester, LongHorn) -> Hyperconvergence
 - CERN
 - BNL
- Автоматизация развёртывания ПО и разработки на основе Gitlab (GitOps, Git Duo, Gitea)
 - CERN
 - BNL

Состояние и будущее дистрибутивов Linux

- **CERN** — <https://indico.cern.ch/event/1477299/contributions/6366311/>
 - AlmaLinux, RedHat, Debian
- **BNL** - <https://indico.cern.ch/event/1477299/contributions/6425485/>
 - AlmaLinux
- **DESY** — смесь RedHat, Debian, другие.
- **KEK** — RedHat
- **RAL** — Rocky9

Problems

- **CERN** - AlmaLinux 9 storage nodes experiencing **XFS** corruption (EOSCMS and EOSLHCb — data corruption and slow restoration)
- Вечная проблема сетевая безопасность -
 - <https://cern.ch/forensics>
 - Создание локальных хранилищ (gitlab.cern.ch)
 - **Zeek** for monitoring and analysis network traffic
- Внимание к экономии ресурсов всякого рода (электроэнергии, воды, прочего) — 6 выступлений (например, <https://indico.cern.ch/event/1477299/contributions/6381818/>).

Прогресс в оборудовании CPU&GPU

- Updates on CPUs, GPUs and AI accelerators
 - <https://indico.cern.ch/event/1477299/contributions/6381816/>

Data Center AI Chip roadmap

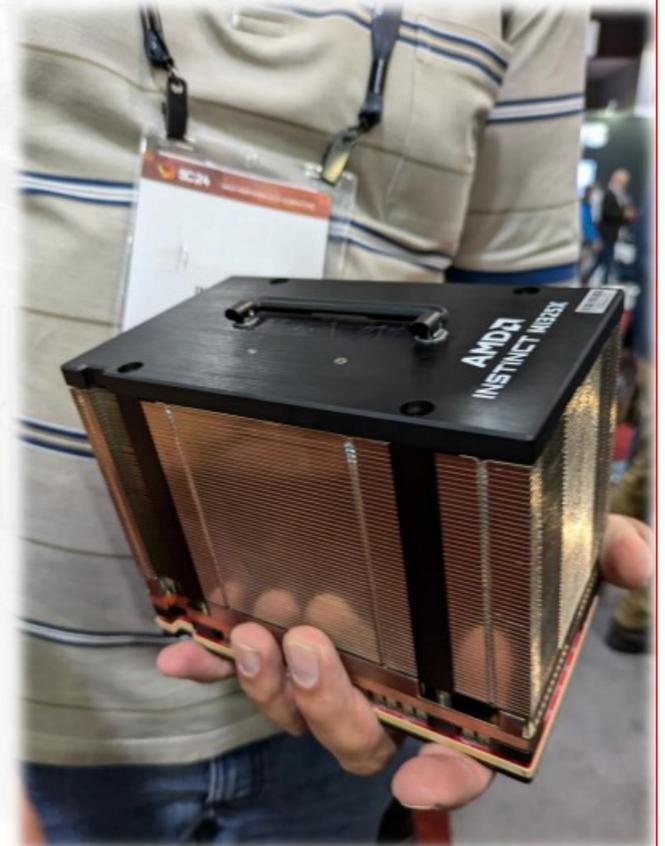
Brand	Name	Node	HBM Generation	HBM Capacity [GByte]	2023		2024			2025			2026			2027		
					Q1	Q2	Q3	Q4	Q1	Q2	Q3	Q4	Q1	Q2	Q3	Q4	Q1	Q2
NVIDIA	A100	N7	HBM2E	80														
	H100	N5	HBM3	96														
	GH200	N5	HBM3	96														
	H20	N5	HBM3	96														
	H200	N5	HBM3E	141														
	B200	N4P	HBM3E	192														
	GB200	N4P	HBM3E	192x2														
	B200A	N4P	HBM3E	96														
	B200 Ultra	N4P	HBM3E	144														
	B20	N4P	HBM3E	96														
	GB200 Ultra	N4P	HBM3E	288														
GB200A	N4P	HBM3E	144															
Rubin	N3	HBM4	384															
Rubin Ultra	N3	HBM4	576															
AMD	MI250X	N6	HBM2E	128														
	MI300X	N5/N6	HBM3	192														
	MI325X	N5/N6	HBM3E	288														
	MI350	N3	HBM3E	288														
	MI400																	
intel	Gaudi2	N7	HBM2E	96														
	Gaudi3	N5	HBM2E	128														
	Falcon Shores	N3																

Cores	AMD EPYC	CCD (Zen5/Zen5c)	Base/Boost* (up to GHz)	Default TDP (W)	L3 Cache (MB)	Price (1 KU, USD)
192 cores	9965	"Zen5c"	2.25 / 3.7	500W	384	\$14,813
160 cores	9845	"Zen5c"	2.1 / 3.7	390W	320	\$13,564
144 cores	9825	"Zen5c"	2.2 / 3.7	390W	384	\$13,006
128 cores	9755	"Zen5"	2.7 / 4.1	500W	512	\$12,984
	9745	"Zen5c"	2.4 / 3.7	400W	256	\$12,141
96 cores	9655	"Zen5"	2.6 / 4.5	400W	384	\$11,852
	9655P	"Zen5"	2.6 / 4.5	400W	384	\$10,811
	9645	"Zen5c"	2.3 / 3.7	320W	384	\$11,048
72 cores	9565	"Zen5"	3.15 / 4.3	400W	384	\$10,486
	9575F	"Zen5"	3.3 / 5.0	400W	256	\$11,791
64 cores	9555	"Zen5"	3.2 / 4.4	360W	256	\$9,826
	9555P	"Zen5"	3.2 / 4.4	360W	256	\$7,983
	9535	"Zen5"	2.4 / 4.3	300W	256	\$8,992
	9475F	"Zen5"	3.65 / 4.8	400W	256	\$7,592
48 cores	9455	"Zen5"	3.15 / 4.4	300W	192	\$5,412
	9455P	"Zen5"	3.15 / 4.4	300W	192	\$4,819
36 cores	9365	"Zen5"	3.4 / 4.3	300W	256	\$4,341
	9375F	"Zen5"	3.8 / 4.8	320W	256	\$5,306
32 cores	9355	"Zen5"	3.55 / 4.4	280W	256	\$3,694
	9355P	"Zen5"	3.55 / 4.4	280W	256	\$2,998
	9335	"Zen5"	3.0 / 4.4	210W	256	\$3,178
24 cores	9275F	"Zen5"	4.1 / 4.8	320W	256	\$3,439
	9255	"Zen5"	3.25 / 4.3	200W	128	\$2,495
16 cores	9175F	"Zen5"	4.2 / 5.0	320W	512	\$4,256
	9135	"Zen5"	3.65 / 4.3	200W	64	\$1,214
	9115	"Zen5"	2.6 / 4.1	125W	64	\$726
8 cores	9015	"Zen5"	3.6 / 4.1	125W	64	\$527

AMD GPU

AMD - GPUs

- **Instinct MI350** series available later in 2025. Based on 3nm node, **added support for FP4/FP6** (like in Blackwell GPUs).
 - Based on CDNA 4 architecture
- **Instinct MI400** will follow MI350, and will be based on «CDNA next» architecture
- CDNA 3 architecture is expected to bring an **8x** increase over CDNA 2
- CDNA 4 architecture is expected to offer around a **35x increase in AI inference performance** over CDNA 3GPUs.
- **Software stack** seems to be the major weak point, compared to NVidia



Ожидаемые крупные источники экспериментальных данных

- HL LHC ~2030
- Square Kilometre Array (SKA) ~2029
- Cherenkov Telescope Array Observatory (CTAO) ~2028
- Facility for Antiproton and Ion Research (FAIR) ~2028

Заключение о готовности к HL LHC

- **Technology tracking essential to make cost-efficient choices for HEP computing**
 - Done in different contexts in our community
 - **Many server hardware components are rising in price due to the AI boom**
 - Memory, GPUs, flash, HDD are all affected
 - **AMD, Arm and Intel show healthy competition**
 - A lot of attention to performance/Watt
 - **Evolution of GPUs is not going in a direction very useful for us**
 - FP32/64 performance not increasing
 - **Shipped storage capacity increasingly driven by the global trend**
 - SSDs, HDDs and tape all still relevant and making progress
 - **Network bandwidth correspondingly increasing on LAN and WAN**
 - To cope with increase in cores and storage/server
 - For LHC, driven by HL-LHC data rates
 - **Sustainability is more important than ever**
 - CO2 emissions, liquid cooling, electricity costs and distribution
-
- HEPiX Technology Watch Working Group Report
 - <https://indico.cern.ch/event/1477299/contributions/6373271/>
 - From the experiment plots, with a flat budget, a 15%/year price reduction is required even in the most optimistic case (for CPU and disk).
 - Situation much better than in the past, but still at the limit!

Популярные большие языковые модели

LLMs in 2025

LLM	Developer	Multimodal?	Reasoning?	Access
GPT-4o	OpenAI	Yes	No	Chatbot and API
o3 and o1	OpenAI	No	Yes	Chatbot and API
Gemini	Google	Yes	No	Chatbot and API
Gemma	Google	No	No	Open
Llama	Meta	No	No	Chatbot and open
R1	DeepSeek	No	Yes	Chatbot, API, and open
V3	DeepSeek	No	No	Chatbot, API, and open
Claude	Anthropic	Yes	Yes	Chatbot and API
Command	Cohere	No	No	API
Nova	Amazon	Yes	No	API
Large 2	Mistral AI	Yes (Pixtral)	No	API
Qwen	Alibaba Cloud	Yes (Qwen2.5-VL)	No	Chatbot, API, and open
Phi	Microsoft	No	No	Open
Grok	xAI	No	Yes	Chatbot and open

From «HEPiX Technology Watch Working Group Report» <https://indico.cern.ch/event/1477299/contributions/6373271/>

One page from Lugano Wrap-up

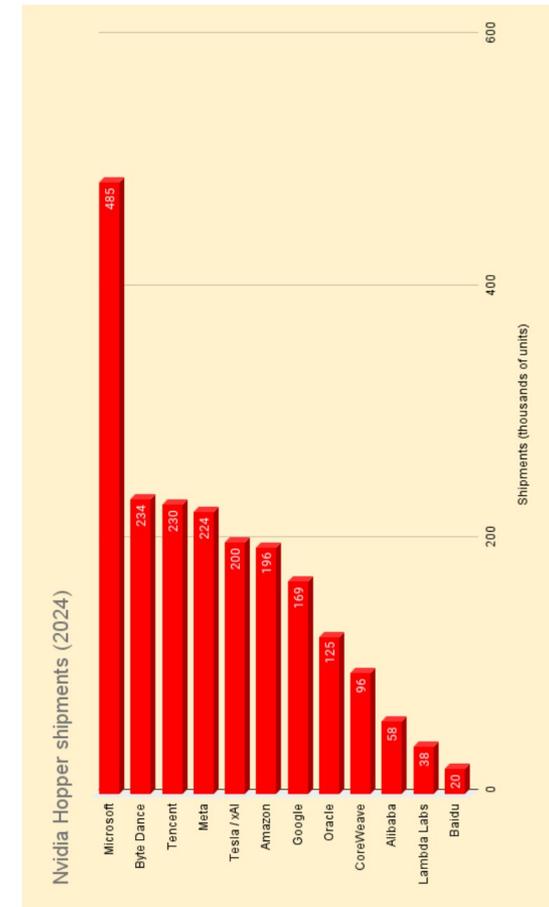
- Ricardo presented the **CERN's Microsoft SQL Server infrastructure**, detailing server management, database solutions, high availability, backup strategies, and data integrity measures for critical applications
- The evolution of **PIC's Big Data Analysis Facility** was shown by Francesc, detailing its Hadoop-based architecture, key services like CosmoHub and SciPIC, integration with PIC's computing ecosystem, and future plans for interoperability, identity federation, and multi-messenger astronomy support
- Andrey showed that **Machine Learning**, particularly Retrieval Augmented Generation (RAG) with a large language model (LLM), aids developers and administrators in **managing complex computing systems**
- Ruben introduced the **Transcription and Translation as a Service (TTaaS)** project at CERN, which leverages Automated Speech Recognition (ASR) and translation technologies to enhance accessibility, diversity, and searchability of CERN's audiovisual content

LLM

Large Language Model (LLM) — тип системы машинного обучения, который используется для задач генерации последовательности на каком-то языке в ответ на ввод запроса на том же или другом языке.

- One natural language to another natural language (NL)
- NL to mathematic/physics formulas
- NL to multimedia object
- etc

NVIDIA GPU распределение поставок in 2024 среди заказчиков



https://regmedia.co.uk/2024/12/20/nvidia_omdia_estimates_2024.png

Наше применение технологии Machine Learning (ML)

- Начальная идея состояла в реализации *искусственного ассистента*, где разработчик (или администратор) задаёт *искусственному ассистенту* вопрос на естественном языке и на нём же получает ответ.
- В качестве первого шага: был разработан *chatbot* для общения с *локальным описанием* сегмента серверной сети.

Выбор архитектуры

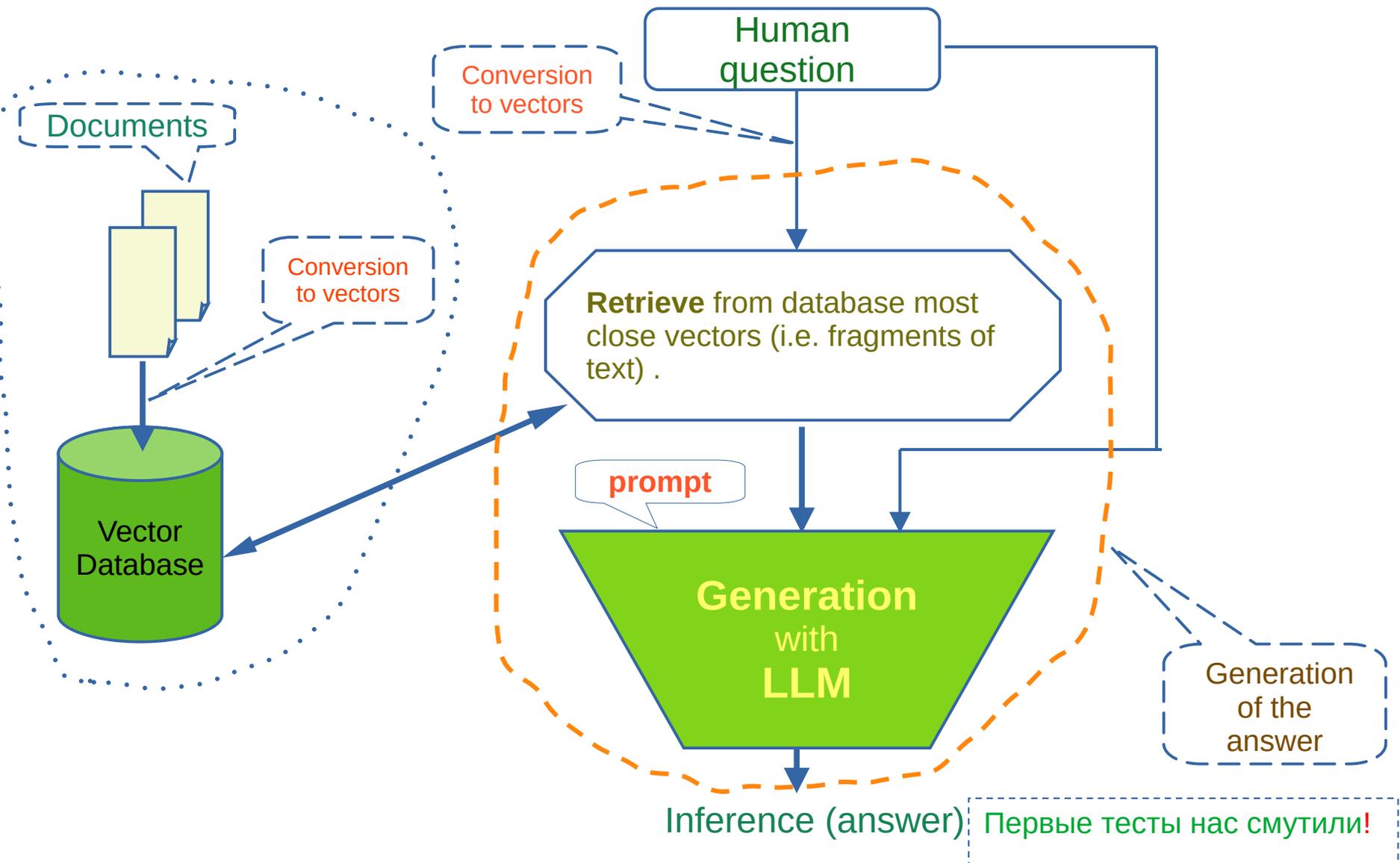
- Была выбрана архитектура Retrieval Augmented Generation (RAG).
- **Retrieval** - в **локальной** базе данных, которая содержит описание **локального** сегмента серверной сети, производится поиск фрагментов текста, которые семантически близки к введённому вопросу.
- **Augmented Generation** означает генерацию с использованием имеющейся pre-trained LLM с использованием **найденных текстовых фрагментов**.
- **Преимущества:**
 - Pre-trained LLM — “передача знаний”.
 - Все операции (поиск, генерация) производятся на **локальном** сервере или кластере, т. е. никакие документы не передаются за пределы локальной сети.
 - Для **локальных** задач разработчиков (или администраторов) не требуются огромные вычислительные ресурсы.

Testbed setup

- The Linux server Intel(R) Xeon(R) CPU E5-2630 v4 @ 2.20GHz, 128GB, four simple GPUs “GeForce GTX 1080 Ti”, python3.12.
- Первоначальный вариант программы на python для архитектуры RAG получена по ссылке <https://gitlab.com/rahasa-k-labs/iollama> (значительно переработана).
- Использовались несколько свободно распространяемых LLM (llama3.2, mixtral, aya:35b, qwen2.5:32b, etc) на платформе ollama.
- Для выполнения **embedding** (т. е. преобразование текста в векторное представление) использовалось несколько свободно распространяемых моделей (LaBSE, paraphrase-multilingual-MiniLM-L12-v2, sbert_large_nlu_ru, etc.) доступных на сайте huggingface.co.
- Фронт-энд был реализован на основе пакета **streamlit** <https://github.com/streamlit>
- Текст описания **локального** сегмента *серверной сети* составлял около 150 страниц.

Basic (naive) Retrieval Augmented Generation (RAG) - emerged in 2021

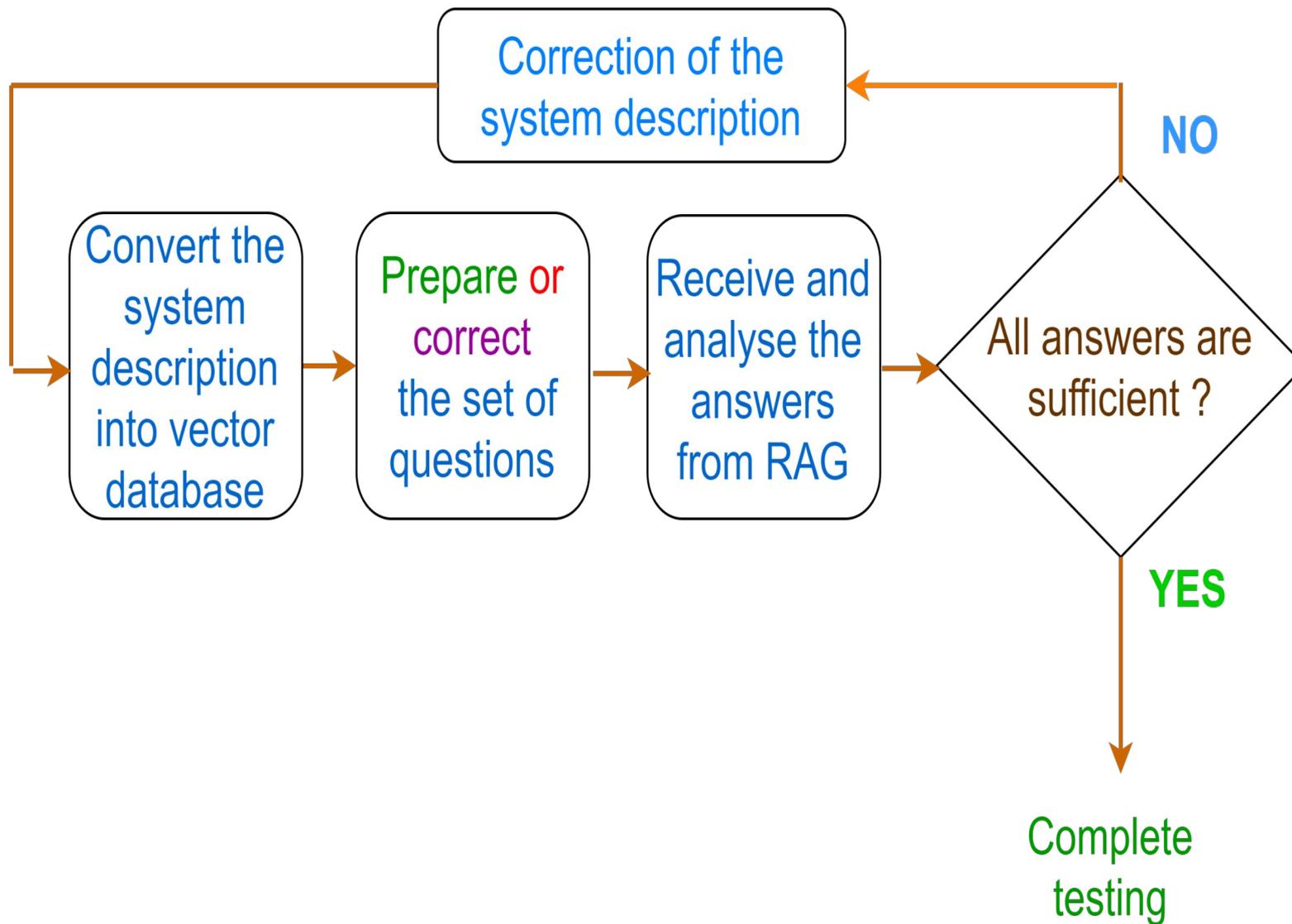
Conversion to vectors = embedding



Начальное тестирование

- Около десяти вопросов были подготовлены и введены один за другим.
- Ответы не выглядели многообещающими:
 - некоторые были полностью неверными,
 - другие были фантазиями (галлюцинациями),
 - часть были не совсем верными.
- Анализ описания сегмента серверной сети показал, что некоторые части описания были не столь ясными и полными, как следовало бы.
 - **Наше решение:** описание системы необходимо улучшить.

Цикл улучшения описания серверного сетевого сегмента



Результат

- В цикле взаимодействия было сделано следующее:
 - Уточнены формулировки и добавлены несколько дополнительных разделов в текстовое описание.
 - Разработчики неоднократно обсуждали что есть что и откуда это берется...
 - Другими словами, архитектура RAG использовалась в качестве технического помощника разработчиков.
- Результат - значительно улучшенная версия описания сегмента серверной сети.

Условия успеха

- Близость тематики **LLM** к тематике документов.
- Близость **embedding** к используемому языку / жаргону.
- Точность инструкций (**prompt**) для модели LLM.
- Полнота и ясность исходного текстового документа.
- **Квалификация команды разработчиков документа.**

Достоинства применения

- Использование архитектуры RAG в разработке или администрировании упрощает/сокращает:
 - Время на разработку описания.
 - Привлечение новых участников разработки / администрирования.
 - Сопровождение разработанной системы.
 - Дальнейшую модернизацию в процессе эксплуатации.
 - Создание цифрового двойника разрабатываемой (или разработанной) системы.
 - Обсуждения внутри команды разработчиков (или администраторов) для прояснения архитектуры RAG создаёт большее взаимопонимание внутри команды.

Пример фронт-энда

The history of your questions and RAG answers

You: 2025-03-19 12:00:25 -- batch processing: which is difference in between 'GE' and 'SGE' ?

RAG: 2025-03-19 12:03:45 -- { "answer": "<think> Okay, I need to explain the differences between Grid Engine (GE) and Sun Grid Engine (SGE) based on the provided context. Let me go through each page to gather relevant information.

On page 34, it mentions that Open Grid Scheduler GE-2011.11p1 is installed on pcfarm-old. It also notes that this version is quite different from the old SGE 5.3 for administrators but remains almost the same for users. The commands for starting and stopping are /etc/init.d/sgemaster and sgeexecd.

Page 40 talks about Sun Grid Engine (SGE) 8.1.10 installed on pcfarm with AlmaLinux 9.3. It uses systemctl commands for service management, which is different from the init scripts used in GE. The configuration directory is also mentioned as /usr/gridengine/default/spool/qmaster/.

Looking at page 10, it explains that GE (Grid Engine) is on pcfarm-sl6 with SL 6.7, while SGE 8.1.10 is on pcfarm with AlmaLinux 9.4. This indicates different operating environments.

High Energy Physics Division

Computing Systems Department:

The platform prototype for developers and administrators to search the answers

in the description for server network for HEPD

=====

RAG_model: deepseek-r1:32b

RAG_embed_model: multilingual-e5

RAG_template: csd_template_eng.txt

RAG_work_dir: /Users/andreyshevel/IOLAAMA-2024-07-25/iollama

RAG_docs: csd

You entered: 2025-03-19 12:00:25 -- batch processing: which is difference in between 'GE' and 'SGE' ?

-- services-notes-2025-03-18.pdf

The answer from the prototype

Please check any answer from RAG!

2025-03-19 12:03:45 -- The RAG answer: { "answer": "<think> Okay, I need to explain the differences between Grid Engine (GE) and Sun Grid Engine (SGE) based on the provided context. Let me go through each page to gather relevant information.

On page 34, it mentions that Open Grid Scheduler GE-2011.11p1 is installed on pcfarm-old. It also notes that this version is quite different from the old SGE 5.3 for administrators but remains almost the same for users. The commands for starting and stopping are `/etc/init.d/sgemaster` and `sgeexecd`.

Page 40 talks about Sun Grid Engine (SGE) 8.1.10 installed on pcfarm with AlmaLinux 9.3. It uses `systemctl` commands for service management, which is different from the init scripts used in GE. The configuration directory is also mentioned as `/usr/gridengine/default/spool/qmaster/`.

Looking at page 10, it explains that GE (Grid Engine) is on pcfarm-sl6 with SL 6.7, while SGE 8.1.10 is on pcfarm with AlmaLinux 9.4. This indicates different operating environments.

From these points, I can outline the differences in versions, OS compatibility, service management commands, and configuration directories between GE and SGE. </think>

The differences between Grid Engine (GE) and Sun Grid Engine (SGE) based on the provided context are as follows:

1. Version Differences:

- **GE:** Version GE-2011.11p1 is installed on pcfarm-old, which is significantly different from the older SGE 5.3 for administrators but remains largely unchanged for users.
- **SGE:** Version 8.1.10 of Sun Grid Engine is installed on pcfarm with AlmaLinux 9.3.

2. Operating System Compatibility:

- **GE:** Runs on Scientific Linux (SL) 6.7 on the server pcfarm-sl6.
- **SGE:** Installed on AlmaLinux 9.4 on the server pcfarm.

3. Service Management Commands:

- **GE:** Uses init scripts for service management, such as `/etc/init.d/sgemaster start` and `/etc/init.d/sgeexecd stop`.
- **SGE:** Utilizes `systemctl` commands, like `systemctl restart autofs`, for managing services.

4. Configuration Directories:

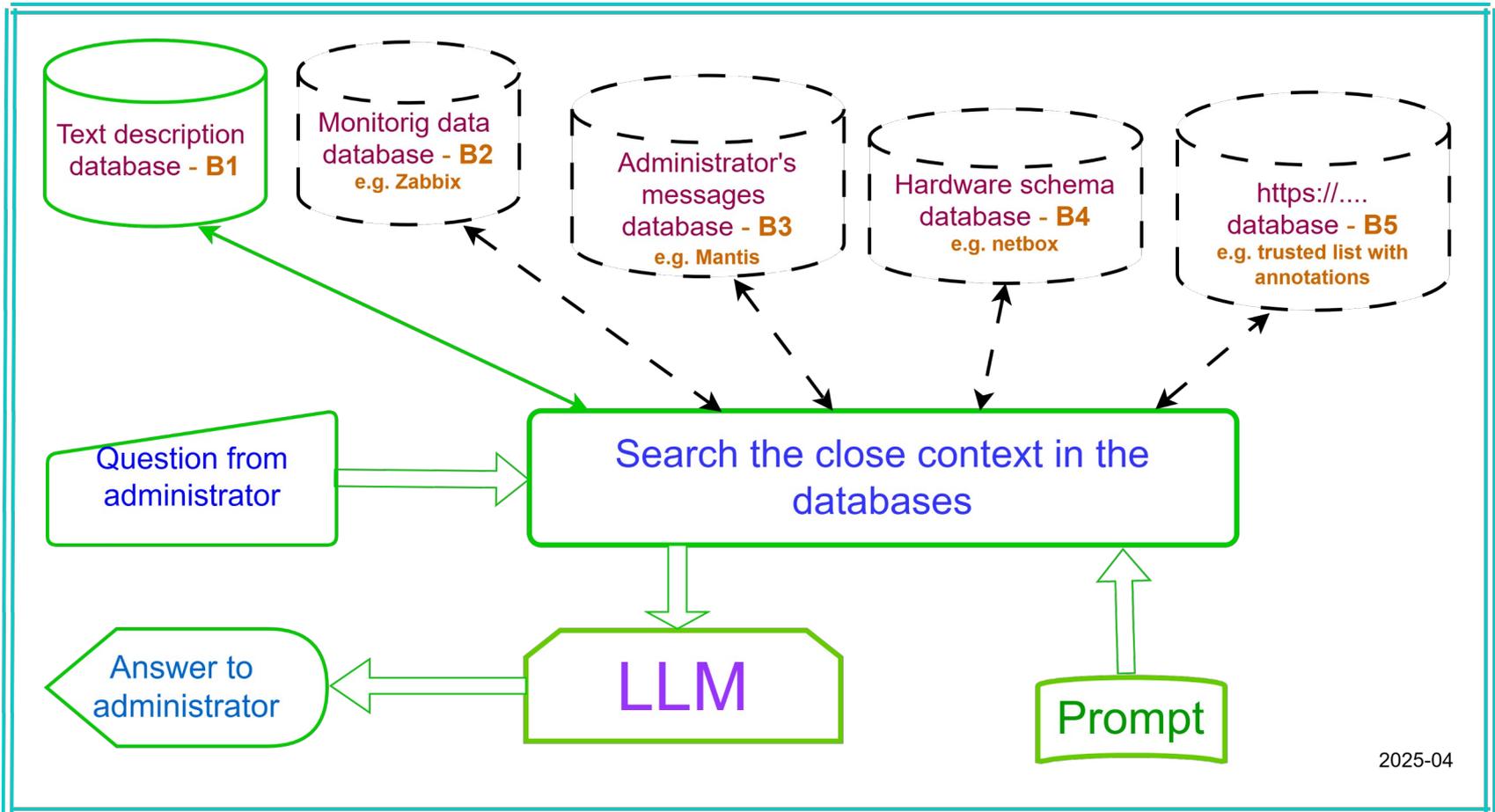
- Both systems have their configuration files in the directory `/usr/gridengine/default/spool/qmaster/`.

In summary, while both GE and SGE serve similar purposes in batch processing, they differ in versioning, operating system support, service management methods, and specific configurations tailored to their respective environments." }

Заключение

- В обозримом будущем описанная технология станет стандартным этапом в создании документации для разрабатываемых систем.
- Разработка архитектуры RAG является неизбежным шагом на пути к созданию полномасштабного цифрового двойника большой разрабатываемой системы.

Дорожная карта дальнейшей разработки



Спасибо за внимание!

References

1. Humza Naveed et al // A Comprehensive Overview of Large Language Models // <https://arxiv.org/pdf/2307.06435.pdf>
2. Shevel, A. // Artificial Neural Networks in High Energy Physics Data Processing (Succinct Survey) and Probable Future Developments. // Phys. Part. Nuclei 55, 309–312 (2024) <https://doi.org/10.1134/S1063779624030742>
3. Alexey Naikov, Anatoly Oreshkin, Alexey Shvetsov, Andrey Shevel // The machine learning platform for developers of large systems // <https://doi.org/10.48550/arXiv.2501.13881> // 2025.
4. Patrick Lewis et al // Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks // <https://arxiv.org/pdf/2005.11401>
5. https://en.wikipedia.org/wiki/Digital_twin

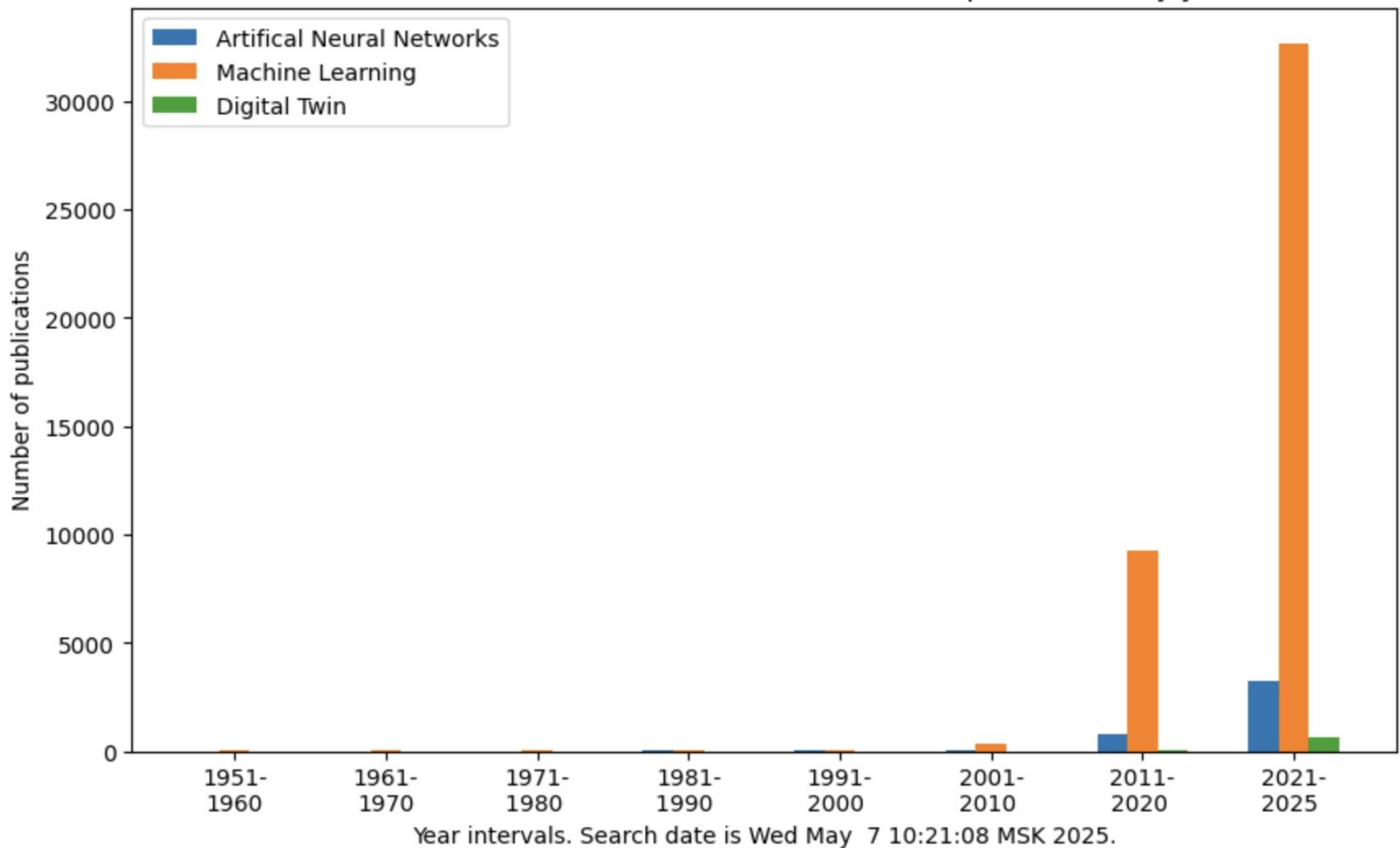
Physics LLM

- Kristian G. Barman et al (22 authors) // Large Physics Models: Towards a collaborative approach with Large Language Models and Foundation Models // <https://arxiv.org/html/2501.05382v1>
- Thorsten Hellert et al // Domain-specific text embedding model for accelerator physics // <https://journals.aps.org/prab/abstract/10.1103/PhysRevAccelBeams.28.044601> (Apr 2025)
- Awesome Scientific Language Models // <https://github.com/yuzhimanhua/Awesome-Scientific-Language-Models>
- Haining Pan et al // Quantum many-body physics calculations with large language models // <https://www.nature.com/articles/s42005-025-01956-y>

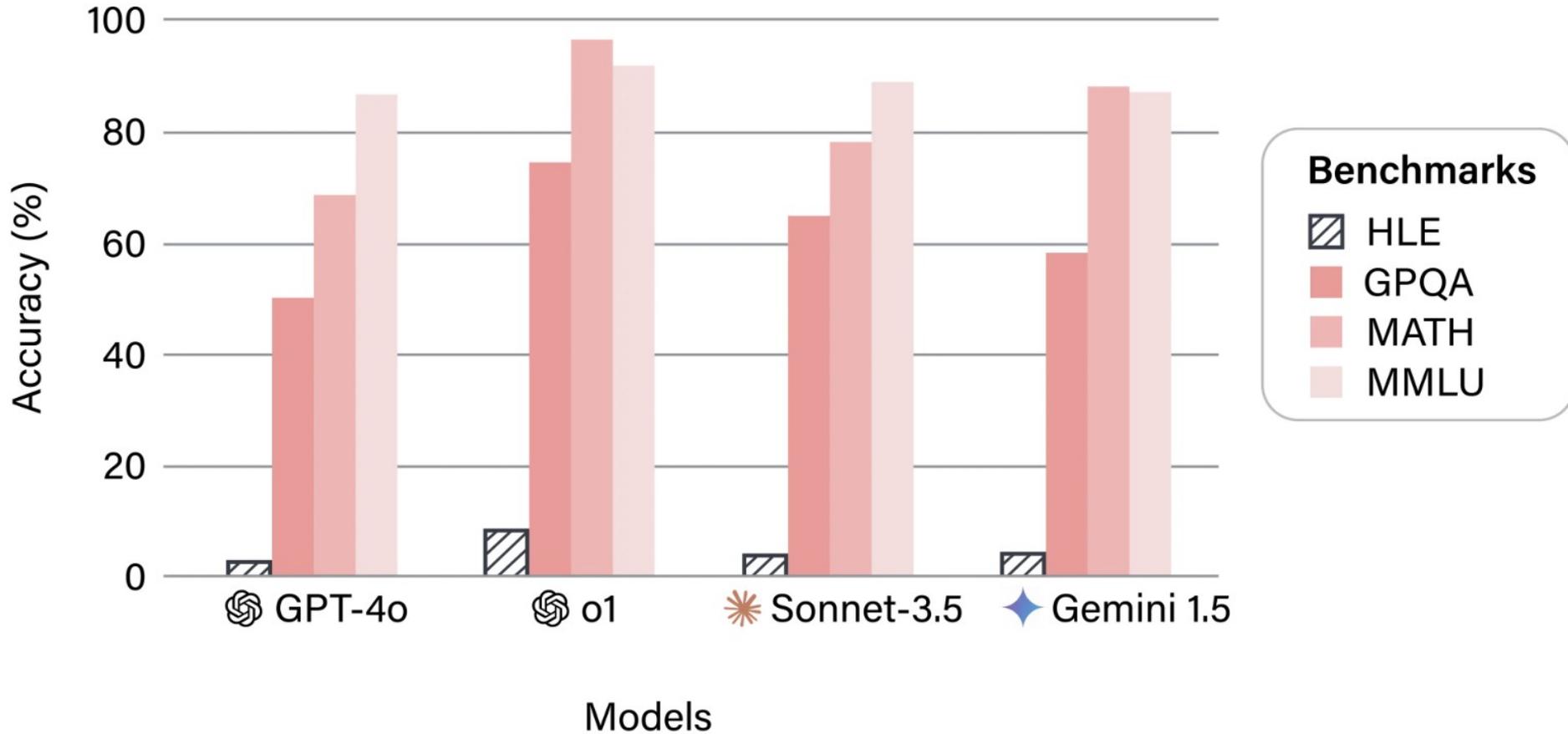
Embedding

- **Embedding** — это преобразование текста документа (описание системы) в векторную базу данных.
- **Embedding** выполняется специальными программными средствами, которые реализованы на основе тренированных искусственных нейронных сетей.

Distribution of mentions of technical terms in Nature.com publications by year interval



Accuracy of LLMs Across Benchmarks



Compared against the saturation of some existing benchmarks, Humanity's Last Exam accuracy remains low across several frontier models, demonstrating its effectiveness for measuring advanced, closed-ended, academic capabilities.

Визит в машинный зал

