MINERVA

# European Support Centre for Scalable AI Research and Deployment

Webinar #1 – MINERVA Services – 2 Dec 2025

Agenda

# Webinar #1 - Agenda

- Welcome and project introduction
- Overview of MINERVA services
- How to contact MINERVA and request support
- Additional activities: benchmarking, best practices
- Q&A session

# Objectives

MINERVA

## MINERVA is a distributed, European-wide HPC-enabled AI application support service

It brings together expertise from EuroHPC Hosting Entities and partners representing major European stakeholders in AI.

MINERVA acts as a central hub for cutting-edge European competences in large-scale ML/AI research and development.

It started in January 2025, and the project's duration is 36 months.

Co-funded by the European Union

EuroHPC Joint Undertaking

This project has received funding from the European High-Performance Computing Joint Undertaking (JU) under grant agreement No 101182737. The JU receives support from the Digital Europe Programme.

# Partners

MINERVA

CINES
Centre Informatique National
de l'Enseignement Supérieur

CINECA

UNIMORE
UNIVERSITÀ DEGLI STUDI DI
MODENA E REGGIO EMILIA

cnrs
INSTITUT DU
DÉVELOPPEMENT ET DES
RESSOURCES EN
INFORMATIQUE
SCIENTIFIQUE

AMD
SILO AI

cea

JÜLICH
Forschungszentrum

Tübingen AI Center

Domyn

BSC
Barcelona
Supercomputing
Center
Centro Nacional de Supercomputación

# What MINERVA Aims To Achieve

## What we do

- Offer a rich service portfolio covering several levels and types of support.
- Interact with AI communities to identify needs and update the MINERVA service portfolio.
- Increase competences of AI communities on AI n HPC.
- Ensure models are developed according to ethical and responsible AI regulations.

## How we do it

- **Support:** Operating an EU-wide support centre.
- **Knowledge transfer**: Publishing best practice guides and guidelines.
- **Benchmarking**: Evaluating model performance on supercomputers.
- **Data Access**: Providing information on access to public datasets.
- **Training Programs**: Providing "AI on HPC" training program.
- **Community Hub**: Supporting large-scale open-source ML/AI research and development on HPC.

# The MINERVA Support

**To deliver support that is both structured and flexible, MINERVA recognises that users may need help on different topics and in different ways.** We organise our services along those two key principles:

**Service Levels (L1–L5)**
These define the thematic areas where support is offered, reflecting the kinds of challenges AI users typically face, from porting workflows to HPC (L1), scaling models and workflows (L2), to pre-training (L3) and fine-tuning foundation models (L4), as well as complying with ethical and regulatory requirements (L5).

**Support Types (T1–T4)**
These describe how the support is delivered – from quick technical help (T1) to long-term engineering collaborations (T4).
📌 *Unlike the Levels, these Support Types are an internal coordination tool to help allocate resources and tailor support formats to user needs.*

# SERVICES

MINERVA

| Support Type | Goal | Scope | Time to resolution | Delivery | Trigger |
|---|---|---|---|---|---|
| **T1 - Assistance** | Quick resolution of specific issues | Debugging, configuration, usage questions | Few days | 🖥️ Remote only | Ticketing system |
| **T2 - Consulting** | Strategic guidance for planning and theoretical questions | Proposals, regulation, architecture, resources | Up to 2 weeks | 🖥️ Remote / 🏢 On-site optional | Ticketing system<br>Scheduled request |
| **T3 - Workshop** | Accelerate workflows through hands-on engineering support | Profiling, performance tuning, scaling | Up to 2 months | 🖥️ Remote / 🏢 On-site optional | Ticketing system<br>Evaluation-based selection & Partner availability |
| **T4 - Embedded Support** | Long-term collaboration on advanced AI projects | Scaling, fine-tuning, large model training, data handling | Up to 6 months | 🖥️ Remote (with meetups recommended) | Ticketing system<br>Evaluation-based selection & Partner availability |

# SERVICES & MORE

MINERVA

**L1**
***Porting AI Applications & Workflows to HPC***

**L2**
***Optimization & Scaling of AI Libraries on HPC***

**L3**
***Pre-training of Large-Scale Foundation Models***

**L4**
***Specialization & Fine-Tuning of Foundation Models***

**L5**
***Ethical & Responsible AI Support***

**Specialised/Advanced trainings for the AI communities**
**ICIAP – EurIPS**

**Catalogues of models and datasets**

**Best Practice Guides on AI on HPC**

**Benchmarks reports**

# Who can access MINERVA support

MINERVA services are open to all actors in the European AI ecosystem – regardless of technical background or experience with HPC. If you're working with AI and face technical, scaling, or regulatory challenges, MINERVA is here to help.

We welcome requests from:

- **Academic researchers** (PhD students, postdocs, principal investigators)
- **Public sector teams** applying AI in areas like health, climate, education, or policy
- **Start-ups and SMEs*** developing AI-driven products or services
- **Larger companies*** with advanced or specialized AI use cases
- **AI engineers and developers** aiming to scale models or optimize performance
- **Project consortia** preparing proposals for national or European funding calls

📌 *No HPC access? We can guide you through the process of applying for resources, choosing the right infrastructure, and getting started with your first project.*

* subject to state-aid regulations

MINERVA

# Minerva Services

Kamel Guerda
MINERVA WP2 Leader

# SERVICES

**L1**
***Porting AI Applications & Workflows to HPC***

**L2**
***Optimization & Scaling of AI Libraries on HPC***

**L3**
***Pre-training of Large-Scale Foundation Models***

**L4**
***Specialization & Fine-Tuning of Foundation Models***

**L5**
***Ethical & Responsible AI Support***

Guidance and Support on **regulations on ethical and responsible AI**

**TYPE 1**
Assistance

**TYPE 2**
Consulting

**TYPE 3**
Workshop

**TYPE 4**
Embedded Support

# T1 – Assistance: First-Line User Support

**Reactive, short interactions to unblock users on concrete issues (jobs, configs, environments, resources).**

- **Example with L1 – Porting**
  ⚠️ Problem: Doctoral researcher gets Slurm errors when launching a training job on an HPC cluster.
  💡 Support: MINERVA reviews the job script and logs to identify misconfigured directives.
  💡 Support: We provide a corrected script and show how to submit and monitor future jobs.

- **Example with L2 – Scaling**
  ⚠️ Problem: SME engineer observes unexpectedly long training times on a multi-GPU setup.
  💡 Support: MINERVA suggests profiling tools to locate bottlenecks in the training workflow.
  💡 Support: We recommend input-pipeline changes to fix data loading and improve GPU utilization.

- **Example with L3 – Specialization**
  ⚠️ Problem: Team wants to fine-tune a language model but is unsure which model and datasets to pick.
  💡 Support: MINERVA leverages ongoing benchmarks and partners' expertise to suggest suitable open models.
  💡 Support: We highlight relevant domain datasets and share references to kick-start the fine-tuning effort.

# T2 – Consulting: Strategic Guidance Sessions

**Structured, planned sessions to provide expert guidance on architecture, scaling, and regulatory questions.**

- **Example with L2 – Scaling**
  ⚠️ Problem: Remote-sensing team wants to scale a high-resolution segmentation model to larger datasets and higher input resolution on HPC.
  💡 Support: MINERVA performs targeted technology watch on distributed training and I/O bottlenecks for similar use cases.
  💡 Support: We give tailored advice on dataset handling, storage layout, and data/model/pipeline parallelism, plus suitable EuroHPC/national resources.

- **Example with L5 – Ethical & Regulatory Compliance**
  ⚠️ Problem: Start-up building clinical decision-support AI needs clarity on how the EU AI Act applies to its use case.
  💡 Support: MINERVA reviews documentation and identifies likely risk classification and applicable obligations.
  💡 Support: We walk through key requirements, share best-practice examples, and provide a checklist for documentation and future audits.

# T3 – Workshop: Hands-On Engineering Sessions

**Intensive 1–2 day hands-on sessions to work directly on users’ code and workflows on HPC systems.**

- **Example with L2 – Optimization**
  ⚠️ Problem: Research lab on satellite object detection faces inefficient training on a GPU cluster.
  💡 Support: MINERVA reviews training scripts and profiling traces ahead of the workshop.
  💡 Support: Together we apply mixed precision, async data loading, and improved GPU scheduling to boost throughput.

- **Example with L3 – Fine-Tuning**
  ⚠️ Problem: Private SME with limited AI experience wants to fine-tune a language model on its internal documentation.
  💡 Support: MINERVA reviews the existing data, infrastructure, and target use cases before the workshop.
  💡 Support: In the session we design the fine-tuning pipeline, apply PEFT (e.g. LoRA), and define evaluation on real internal queries.

# T4 – Embedded Support: Medium- to Long-Term Collaboration

**Embedded collaboration over several months, with a MINERVA engineer working inside the team on development, optimisation, and research around complex AI workflows.**

- Example with L1 – Workflow Adaptation

  ⚠️ Problem: Research group needs to migrate a CPU-based AI workflow to a GPU-accelerated HPC system.

  💡 Support: A MINERVA engineer is embedded for three months to review the codebase and refactor the workflow for GPU execution.

  💡 Support: Together we build a reproducible pipeline with experiment versioning, leading to a stable, documented HPC setup.

- Example with L3 – Pre-training

  ⚠️ Problem: Public research institution wants to pre-train a legal-domain LLM on EuroHPC resources.

  💡 Support: An embedded MINERVA engineer coordinates access, advises on model parallelism, and helps curate and prepare the training dataset.

  💡 Support: We integrate logging/monitoring, ensure compliance with EU data-handling rules, and transfer best practices from MINERVA benchmarks.

# A typical user journey

How a user can move from first access to HPC to advanced collaboration.

1. First contact – Getting onto HPC (T1 Assistance)
   A bioinformatics postdoc contacts MINERVA to access EuroHPC for the first time. Support helps with resource application, environment setup, and first job submissions.

2. Second contact – Becoming efficient (Docs, training, T3 Workshop)
   They use MINERVA guides and training to work more autonomously. When performance issues appear, a T3 Workshop is organised to profile the code and optimise GPU and data loading.

3. Third contact – Planning and scaling (T2 Consulting)
   As the project grows, they request T2 Consulting to discuss scaling to larger datasets, storage and reproducibility, and relevant EuroHPC calls, leaving with a concise roadmap.

# How to Get Support from MINERVA

- For support:
  - minerva4ai.eu/support/#support-form
  - ***support@minerva4ai.eu***

- A unified support team will bind expertise and resource overview across all the partners of the project which will act as common main entry for all the users.

- A multi-level, helpdesk support service will have a single point of entry for users, the AISC portal, backed up by a ticketing tool enabling user requests to be processed and the life of a ticket to be tracked through to resolution, while retaining its history.

- The service portal will be integrated with the ticketing system.

# Services tools and Benchmarking

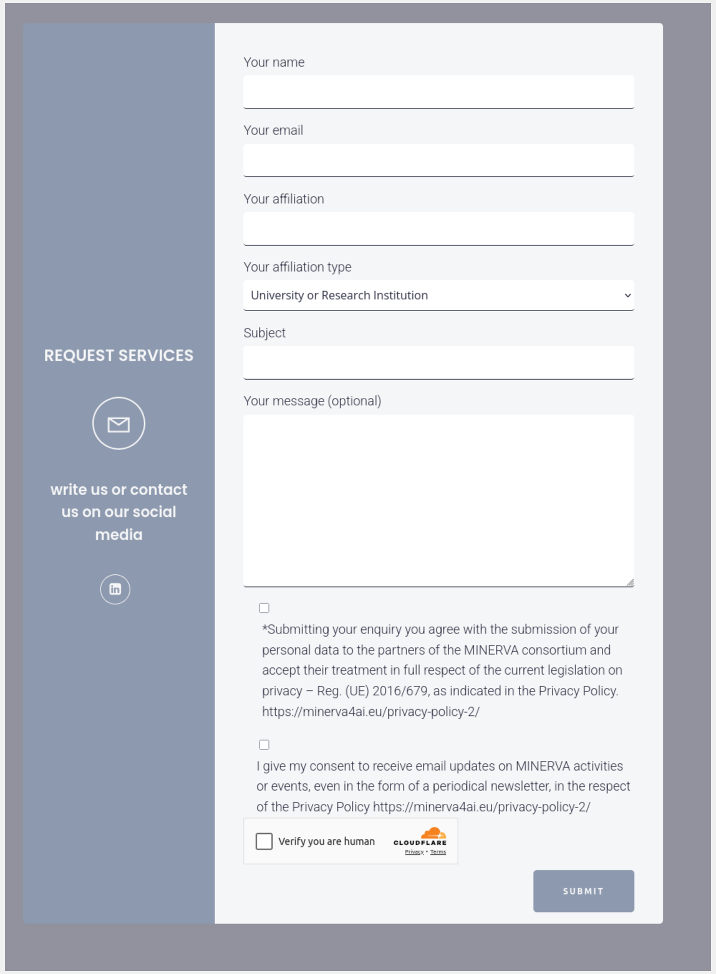
David Vicente
MINERVA WP3 Leader

# MINERVA services for AI support

- Implementation of a ticketing system connected to all the Minerva partners and other external entities providing support to AI communities in Europe.
- Generate a collection of benchmarks to measure the performance of AI machines for specific tasks like training and inference .
- Run the benchmarks in the AI systems of EuroHPC to compare hardware and software.
- Generation of best practice guides linked to AI topics

# User support – ticketing system

https://minerva4ai.eu/support/#support-form



REQUEST SERVICES

write us or contact us on our social media

Your name

Your email

Your affiliation

Your affiliation type

University or Research Institution

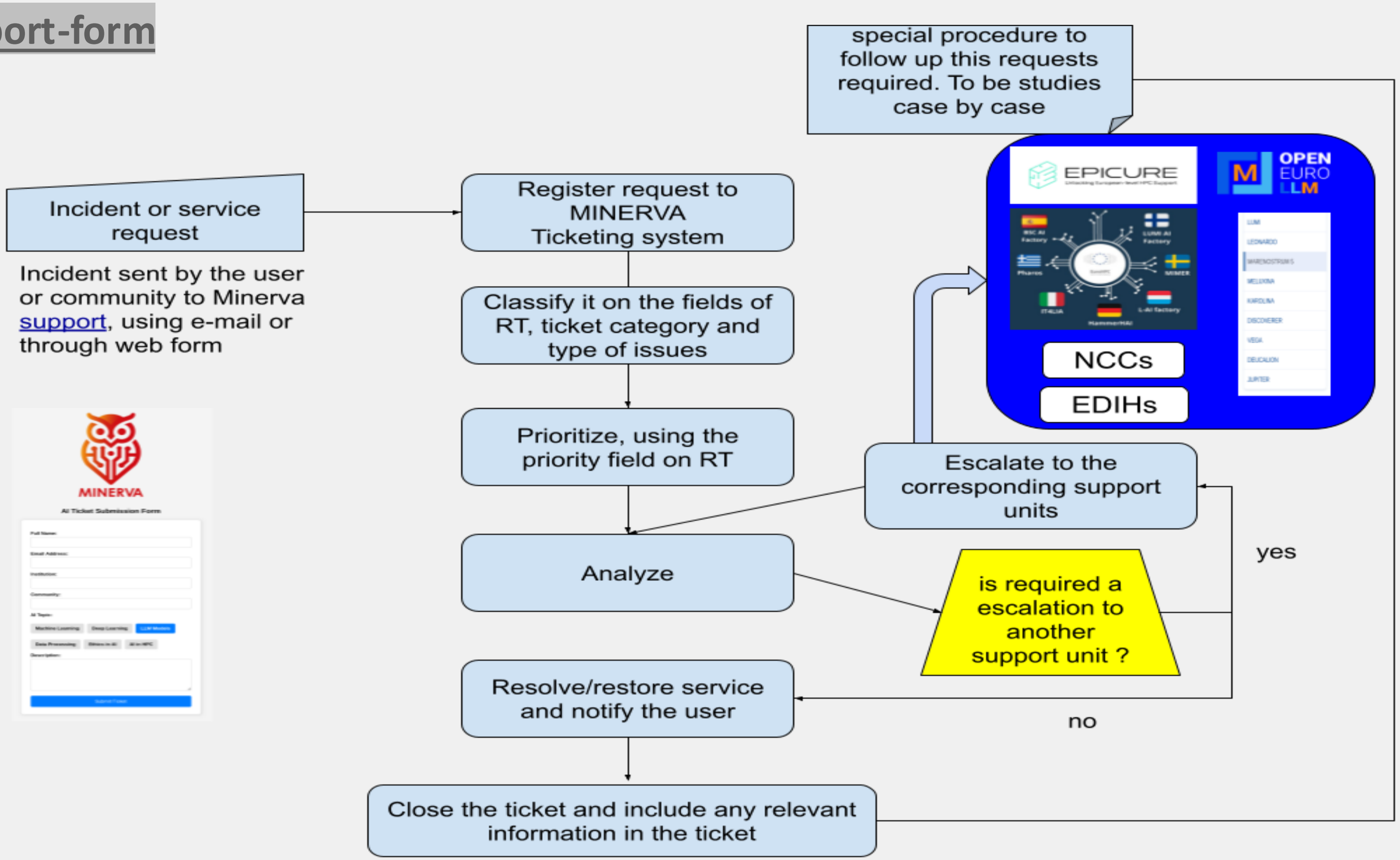
Subject

Your message (optional)

□ *Submitting your enquiry you agree with the submission of your personal data to the partners of the MINERVA consortium and accept their treatment in full respect of the current legislation on privacy – Reg. (UE) 2016/679, as indicated in the Privacy Policy. https://minerva4ai.eu/privacy-policy-2/

□ I give my consent to receive email updates on MINERVA activities or events, even in the form of a periodical newsletter, in the respect of the Privacy Policy https://minerva4ai.eu/privacy-policy-2/

□ Verify you are human



Incident or service request

Incident sent by the user or community to Minerva support, using e-mail or through web form

MINERVA

AI Ticket Submission Form

Full Name:

Email Address:

Institution:

Community:

AI Topic:

Description:

Register request to MINERVA Ticketing system

Classify it on the fields of RT, ticket category and type of issues

Prioritize, using the priority field on RT

Analyze

is required a escalation to another support unit ?

yes

no

Escalate to the corresponding support units

special procedure to follow up this requests required. To be studies case by case

NCCs

EDIHs

Resolve/restore service and notify the user

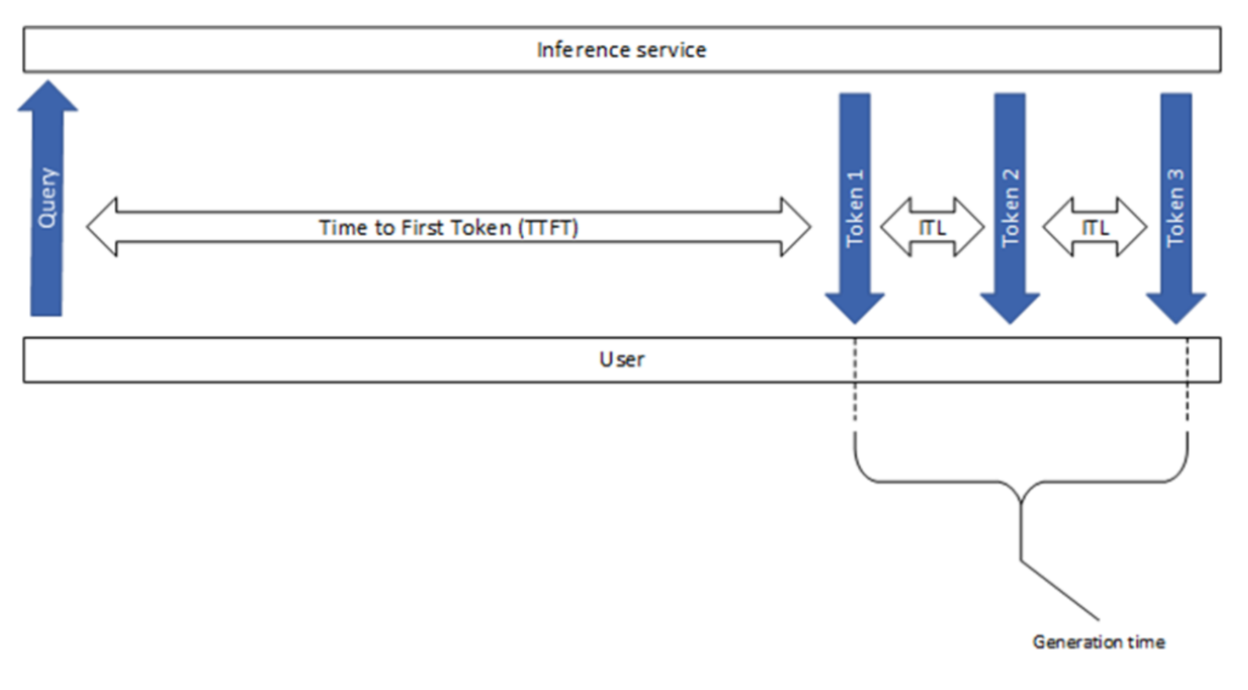
Close the ticket and include any relevant information in the ticket

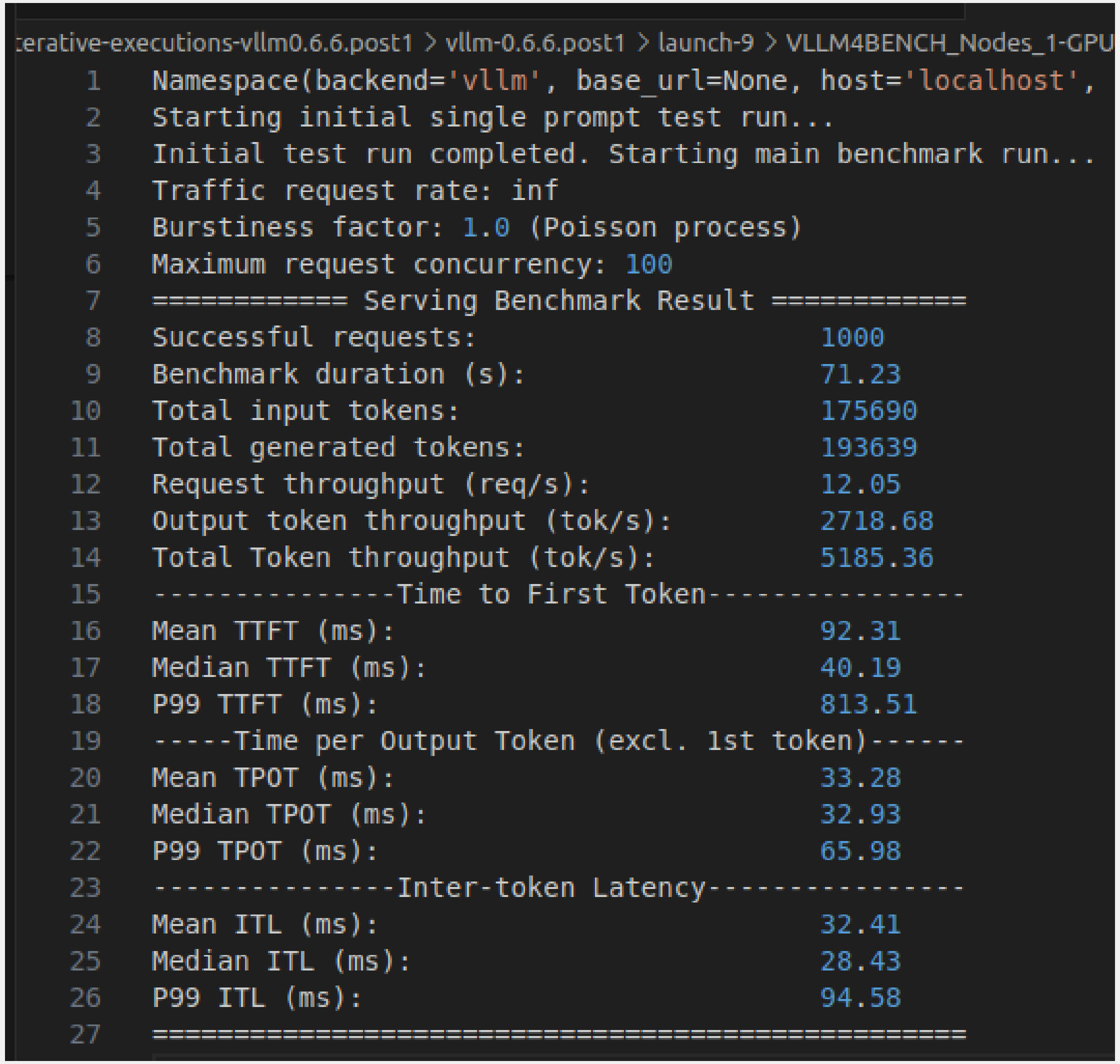
# Benchmarking and Performance Analysis - Objectives

- Benchmark large AI models during both training and inference using diverse datasets and precision settings.
- Evaluate performance on EuroHPC supercomputers across model sizes and frameworks.
- Develop tools and examples for instrumenting and analyzing foundation model computation.
- Support best practices by contributing to guides and public documentation.
- Utilize consortium-developed and external models as use cases for benchmarking.

# Benchmarking Inference - Initial Iteration

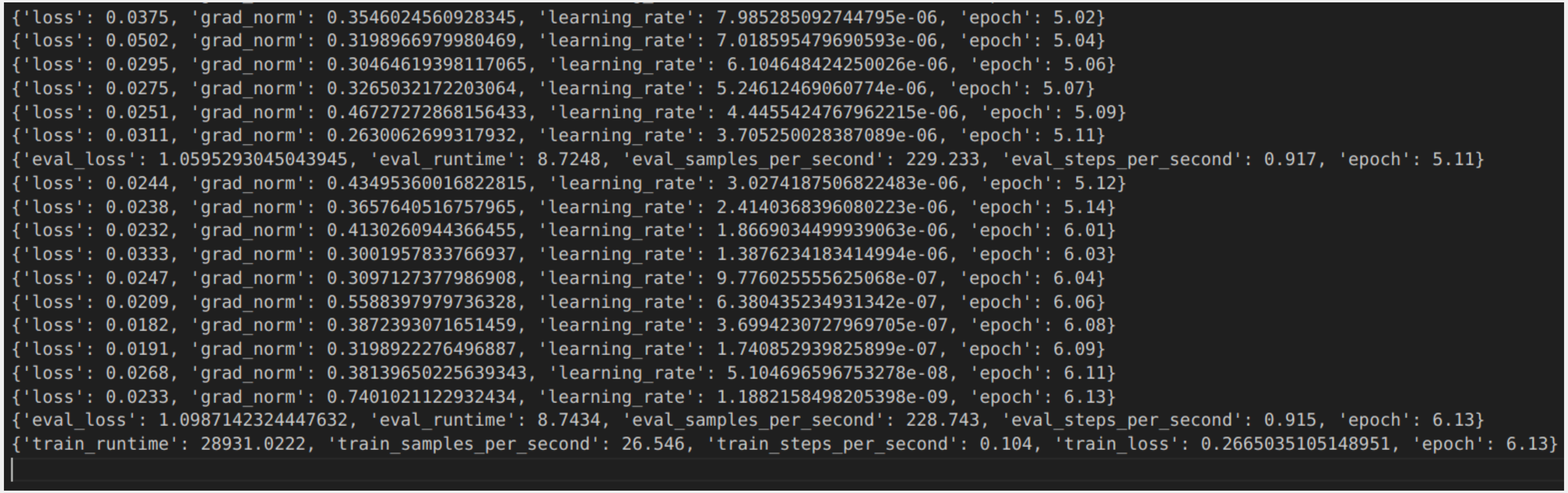
- Conda Environments
- Easy to Use Frameworks:
  - vllm
  - Deepspeed-MII
  - ...
- Examples of Inference




```
:erative-executions-vllm0.6.6.post1 > vllm-0.6.6.post1 > launch-9 > VLLM4BENCH_Nodes_1-GPU
Namespace(backend='vllm', base_url=None, host='localhost',
Starting initial single prompt test run...
Initial test run completed. Starting main benchmark run...
Traffic request rate: inf
Burstiness factor: 1.0 (Poisson process)
Maximum request concurrency: 100
============ Serving Benchmark Result ============
Successful requests:                     1000
Benchmark duration (s):                  71.23
Total input tokens:                      175690
Total generated tokens:                  193639
Request throughput (req/s):              12.05
Output token throughput (tok/s):         2718.68
Total Token throughput (tok/s):          5185.36
---------------Time to First Token----------------
Mean TTFT (ms):                          92.31
Median TTFT (ms):                        40.19
P99 TTFT (ms):                           813.51
-----Time per Output Token (excl. 1st token)------
Mean TPOT (ms):                          33.28
Median TPOT (ms):                        32.93
P99 TPOT (ms):                           65.98
---------------Inter-token Latency----------------
Mean ITL (ms):                           32.41
Median ITL (ms):                         28.43
P99 ITL (ms):                            94.58
==================================================
```

# Benchmarking Training - Initial Iteration

- Conda Environments
- Easy to Use Frameworks:
  - pytorch
  - accelerate
  - deepspeed
- Examples of Training



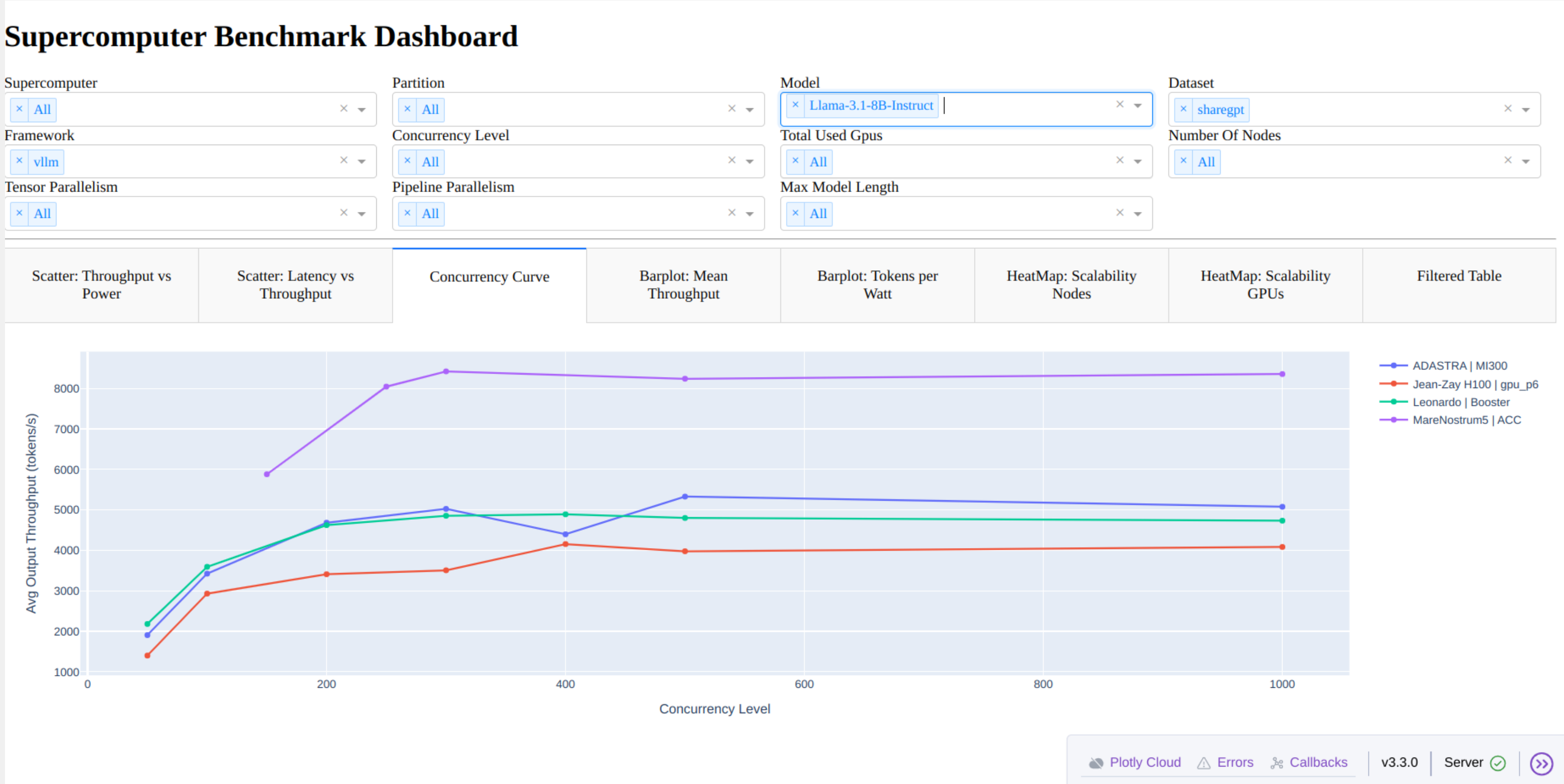
```
{'loss': 0.0375, 'grad_norm': 0.3546024560928345, 'learning_rate': 7.985285092744795e-06, 'epoch': 5.02}
{'loss': 0.0502, 'grad_norm': 0.3198966979980469, 'learning_rate': 7.018595479690593e-06, 'epoch': 5.04}
{'loss': 0.0295, 'grad_norm': 0.30464619398117065, 'learning_rate': 6.104648424250026e-06, 'epoch': 5.06}
{'loss': 0.0275, 'grad_norm': 0.3265032172203064, 'learning_rate': 5.24612469060774e-06, 'epoch': 5.07}
{'loss': 0.0251, 'grad_norm': 0.46727272868156433, 'learning_rate': 4.4455424767962215e-06, 'epoch': 5.09}
{'loss': 0.0311, 'grad_norm': 0.2630062699317932, 'learning_rate': 3.705250028387089e-06, 'epoch': 5.11}
{'eval_loss': 1.0595293045043945, 'eval_runtime': 8.7248, 'eval_samples_per_second': 229.233, 'eval_steps_per_second': 0.917, 'epoch': 5.11}
{'loss': 0.0244, 'grad_norm': 0.43495360016822815, 'learning_rate': 3.0274187506822483e-06, 'epoch': 5.12}
{'loss': 0.0238, 'grad_norm': 0.3657640516757965, 'learning_rate': 2.4140368396080223e-06, 'epoch': 5.14}
{'loss': 0.0232, 'grad_norm': 0.4130260944366455, 'learning_rate': 1.8669034499939063e-06, 'epoch': 6.01}
{'loss': 0.0333, 'grad_norm': 0.3001957833766937, 'learning_rate': 1.3876234183414994e-06, 'epoch': 6.03}
{'loss': 0.0247, 'grad_norm': 0.3097127377986908, 'learning_rate': 9.776025555625068e-07, 'epoch': 6.04}
{'loss': 0.0209, 'grad_norm': 0.5588397979736328, 'learning_rate': 6.380435234931342e-07, 'epoch': 6.06}
{'loss': 0.0182, 'grad_norm': 0.3872393071651459, 'learning_rate': 3.6994230727969705e-07, 'epoch': 6.08}
{'loss': 0.0191, 'grad_norm': 0.3198922276496887, 'learning_rate': 1.740852939825899e-07, 'epoch': 6.09}
{'loss': 0.0268, 'grad_norm': 0.38139650225639343, 'learning_rate': 5.104696596753278e-08, 'epoch': 6.11}
{'loss': 0.0233, 'grad_norm': 0.7401021122932434, 'learning_rate': 1.1882158498205398e-09, 'epoch': 6.13}
{'eval_loss': 1.0987142324447632, 'eval_runtime': 8.7434, 'eval_samples_per_second': 228.743, 'eval_steps_per_second': 0.915, 'epoch': 6.13}
{'train_runtime': 28931.0222, 'train_samples_per_second': 26.546, 'train_steps_per_second': 0.104, 'train_loss': 0.2665035105148951, 'epoch': 6.13}
```

# Benchmarking Training - Next Iterations

## 🔬 Benchmark Expansion

- Integrate standardized benchmarks:
  - MilaBench – Advanced LLM Training benchmarking suite.
  - MLPerf – Industry-standard ML performance benchmarks for Inference/Training.
  - Others
- More diverse workloads:
  - Reinforcement Learning.
  - Prioritize emerging European models and datasets for upcoming benchmarks.
- Public Release & Collaboration:
  - Publish results on MINERVA github (https://github.com/minerva4ai-eu/minerva-benchmarks/).
  - Open-source selected tools and configs for community reproducibility.
  - Prepare materials for Best Practices documentation.

# MINERVA – Benchmarking results



## Supercomputer Benchmark Dashboard

Supercomputer: × All

Partition: × All

Model: × Llama-3.1-8B-Instruct

Dataset: × sharegpt

Framework: × vllm

Concurrency Level: × All

Total Used Gpus: × All

Number Of Nodes: × All

Tensor Parallelism: × All

Pipeline Parallelism: × All

Max Model Length: × All

## Output filter features:

**Dynamic Filters:** Supercomputer, Frameworks, Models, Datasets, Parallelism Type, # GPUs, # Nodes.

**Inference:**

- Throughput vs Concurrency Level
- Latency vs Throughput
- Scalability Heatmaps

**Training/Fine-Tuning:**

- Throughput Scaling
- GPU efficiency
- Total Training Time comparison

## Example:

- **Framework:** vllm
- **Model:** Llama-3.1-8B-Instruct
- **Dataset:** sharegpt
- Across distinct levels of concurrency
- **Systems:** ADASTRA, JeanZay, MN5, Leonardo, JUPYTER (when available)
- Data isn't aligned yet

# Best Practice Guidelines – 2 collections

*Soon available on MINERVA website*

## "Optimizing AI Models on European Infrastructure"

- 📕 *“Infrastructure Foundations for Scalable LLM Training”*
- 📗 *"Efficient and Reliable LLM Pretraining"*
- 📘 *“Post-Training: Alignment and User-Centered Objectives”*
- 📙 *“Explainable AI”*

**Status**: Writing
**Publication**: June 2026

## "Dataset Management for European AI Ecosystems"

- 📕 *“Dataset Sourcing: Where and How to Collect Data”*
- 📗 *“Cleaning, Preprocessing, and Curation”*
- 📘 *“Quality, Ethics, and Compliance”*
- 📙 *"Publishing, Reuse, and Maintenance"*

**Status**: Writing
**Publication**: December 2027

# Q&A

Questions?

LinkedIn

Survey

Reach out to us @ *info@minerva4ai.eu*

# Thank you

Co-funded by the European Union

EuroHPC Joint Undertaking

This project has received funding from the European High-Performance Computing Joint Undertaking (JU) under grant agreement No 101182737. The JU receives support from the Digital Europe Programme.