



High Performance Computing

Andrej Filipcic

Krvavec, 29.9.2026

SLAIF & SLING & HPC

- SLING – Slovenian Consortium for High-Performance Computing
 - 23 members, 4 from industry
 - Strategy for computing evolution in Slovenia
- SLAIF – Slovenian AI Factory, Horizon Europe funded projects
 - 10 partners
 - 3 years (2025-2028)
- AI-HPC (no name yet) – next generation HPC replacing Vega in 2028



Computing Resources in Slovenia

- Supporting HEP Experiments: CPU, disk pledge + opportunistic usage
 - ATLAS, Pierre Auger, Belle 2, LHCb
 - Vera Rubin (LSST)
- Hardware:
 - F9 SiGNET cluster
 - CMI JSI cluster + Ceph (shared with F9, E8)
 - ARNES HPC
 - Vega HPC
 - ~~Maister HPC~~
 - Trdina HPC
- External resources
 - Leonardo – Cineca: part of CPU and GPU partition dedicated to Slovenia
 - IT4LIA – IT AI Factory – co-ownership of HPC
 - AI:AT – Austrian AI Factory – tight collaboration, exchange of resources
 - EuroHPC machines:
 - Through EuroHPC calls for computing projects

Projects

- Active:
 - SLAIF – Slovenian AI Factory
 - EPICURE – high level support for EuroHPC users
 - DaFab – AI Factory for Copernicus Data at Scale
 - EuroCC – EuroHPC Competence Centre
- Past:
 - interTwin – Co-designing and prototyping an interdisciplinary Digital Twin Engine.

Resources available to us

- CPU: (used on average)
 - ~40k cores (80k HT) on Vega
 - 10k cores on ARNES
 - 1500 cores on CMI
 - 3000 cores on SiGNET
- DISK:
 - 7PB on Vega
 - 6PB on CMI
 - +additional short term space
- Pledge:
 - ATLAS: 100k HS06, 12PB disk
 - Belle 2: 10k HS06, 1PB disk
- GPU (non dedicated)
 - 240 A100 40MB on Vega
 - 20 H100 on Arnes (temporary out)
 - FRI: FRIDA B200s and older mix

CMI & F9 resources are 6 years old or more, no plans for refresh

The EuroHPC Supercomputing Ecosystem



EXASCALE



PRE-EXASCALE



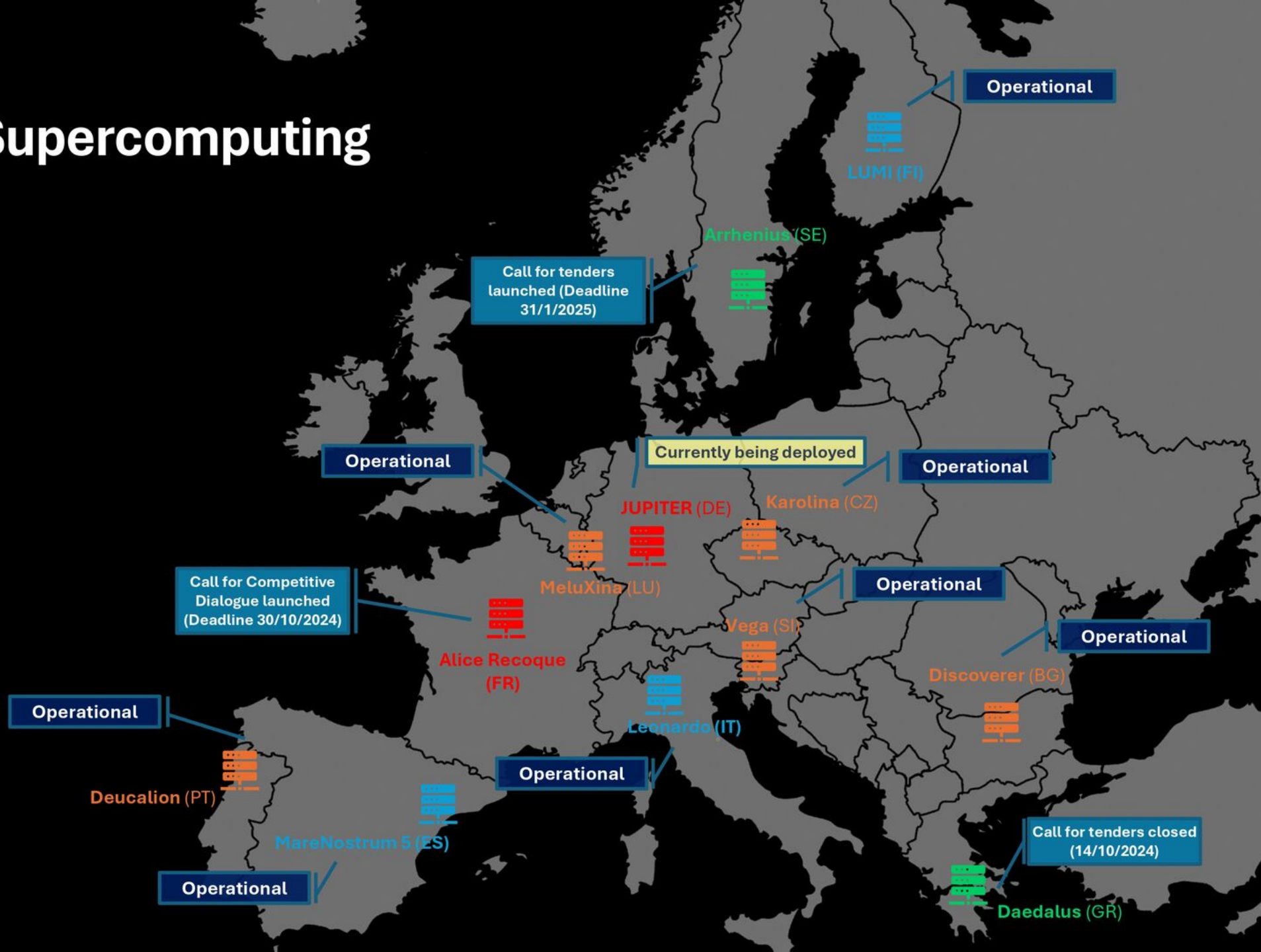
PETASCALE



MID-RANGE

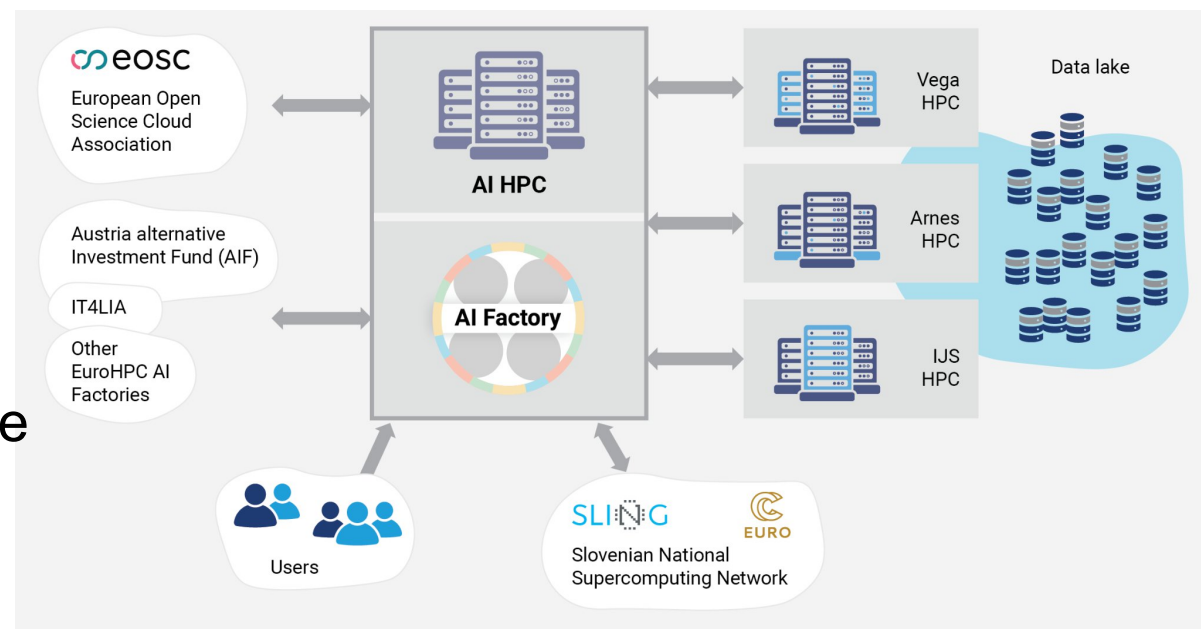


EuroHPC
Joint Undertaking



AI and the AI Factory

- AI: from a basic research area to a general tool (neural net based large generative models)
- European landscape:
 - ~ From research and knowledge basis
 - ~ To regulation and confidence building
 - ~ Finally: infrastructure (*sovereignty*)
- AI Factory: a project an AI-optimized machine **in the future**
- Most people expect **LLMs** but large models demand large infrastructure
 - ~ Data sets
 - ~ Compute power
 - ~ Talent and knowledge



What is an AI Factory?

- **Short answer: AI infrastructure**
in the context of EU independence, sovereignty and global power
 - Where in the 20th centuries we built road and power networks
 - We now build data networks, HPC centres and AI factories
- Components: **physical** (compute, storage), **digital** (data, software) and **human** (knowledge, services) infrastructure.
- Slovenian AI factories is allowing us:
to be more than an user of AI tools, but a developer and participant in EU technology sovereignty and global competitive power: to work with our neighbours and participate in the EuroHPC landscape.



Why do we need and AI factory?

To lower the barrier from UI experiment to production

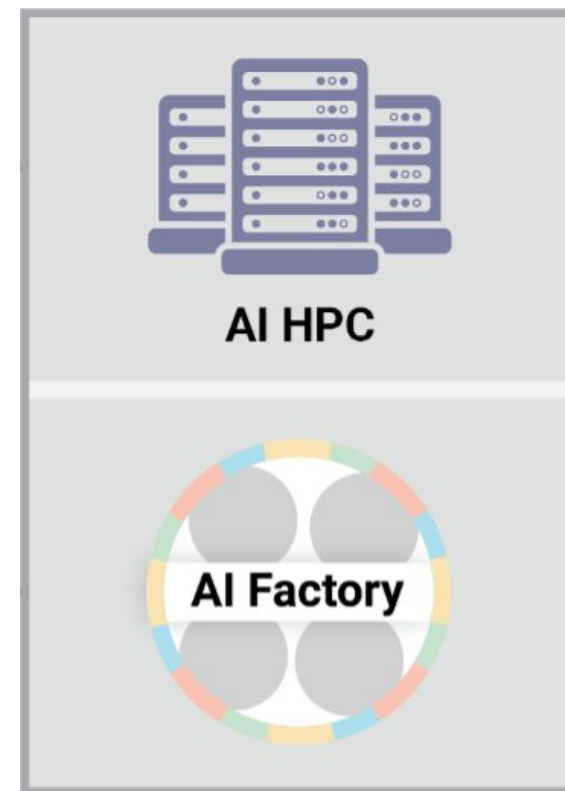
- A complete set of services for companies and researcher to support development and deployment of European (generative) UI models and tools

Two main AIF components:

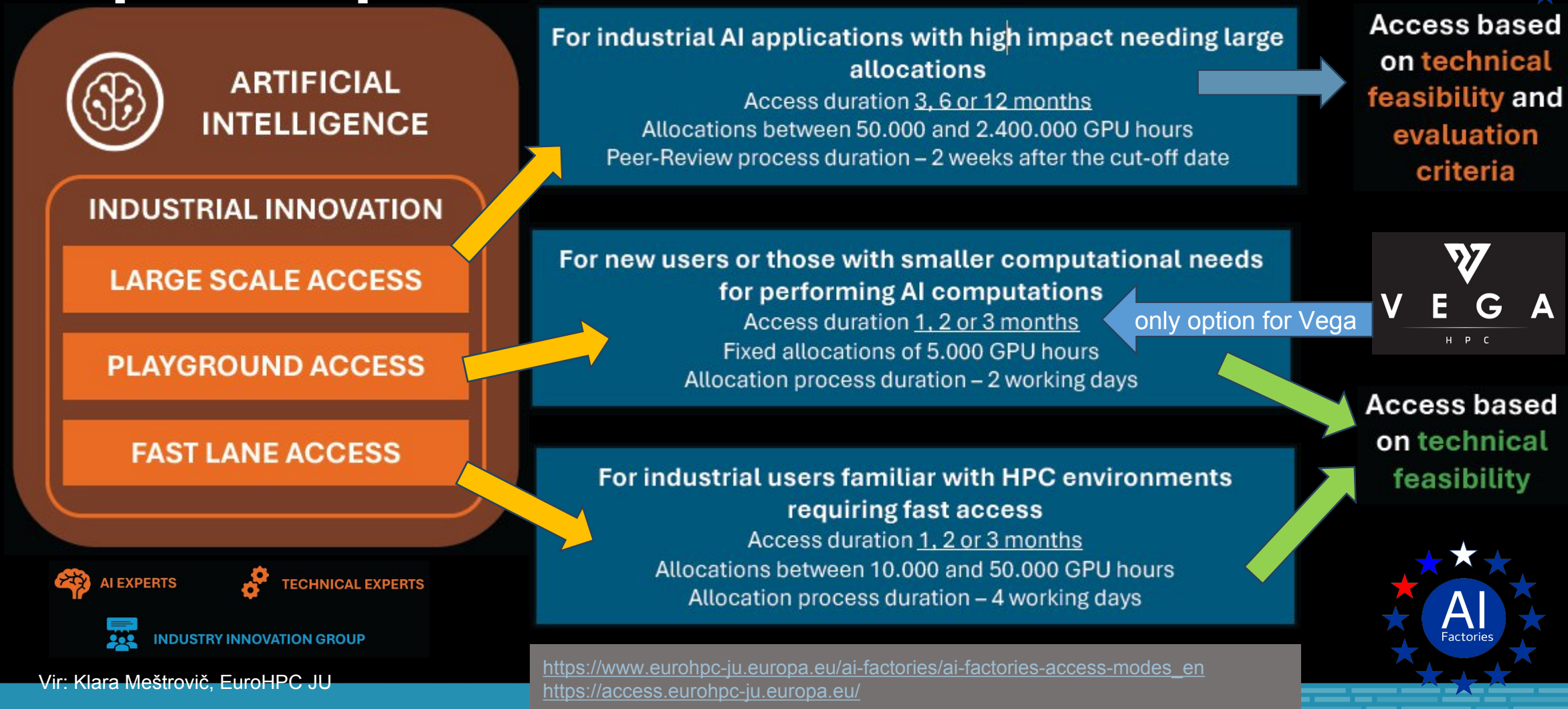
- An AI-optimized supercomputer
- AI Factory services and facilities

AI Factories combine:

- Computer power
- Algorithms and workflows
- Data sets and AI models
- Talent, knowledge, skills and services

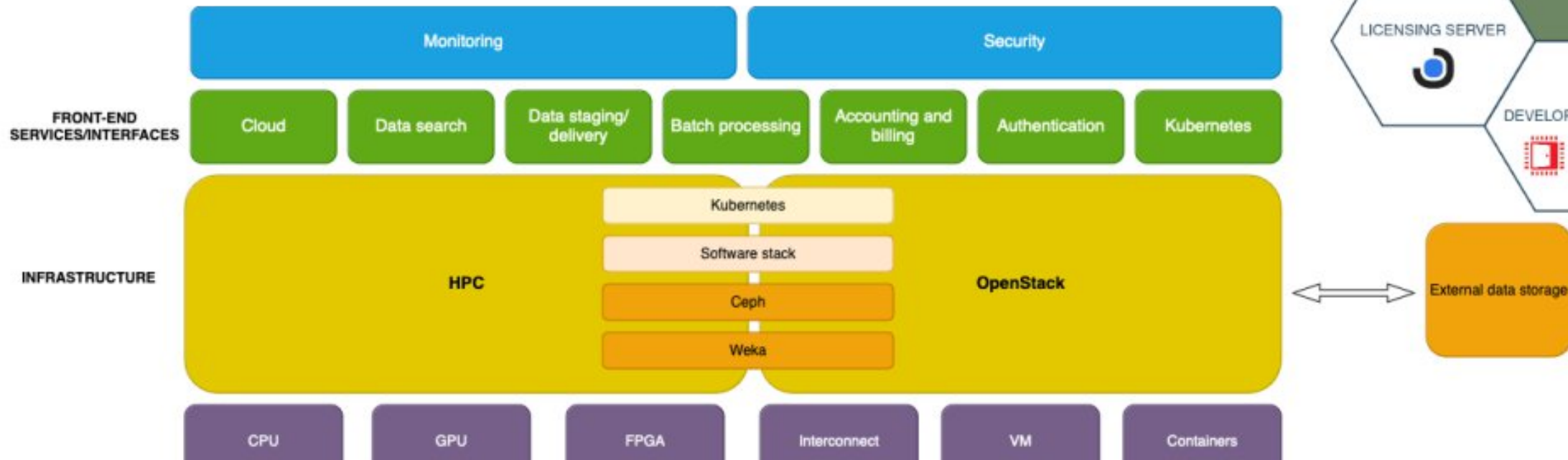


Industry access to EuroHPC AI supercomputers

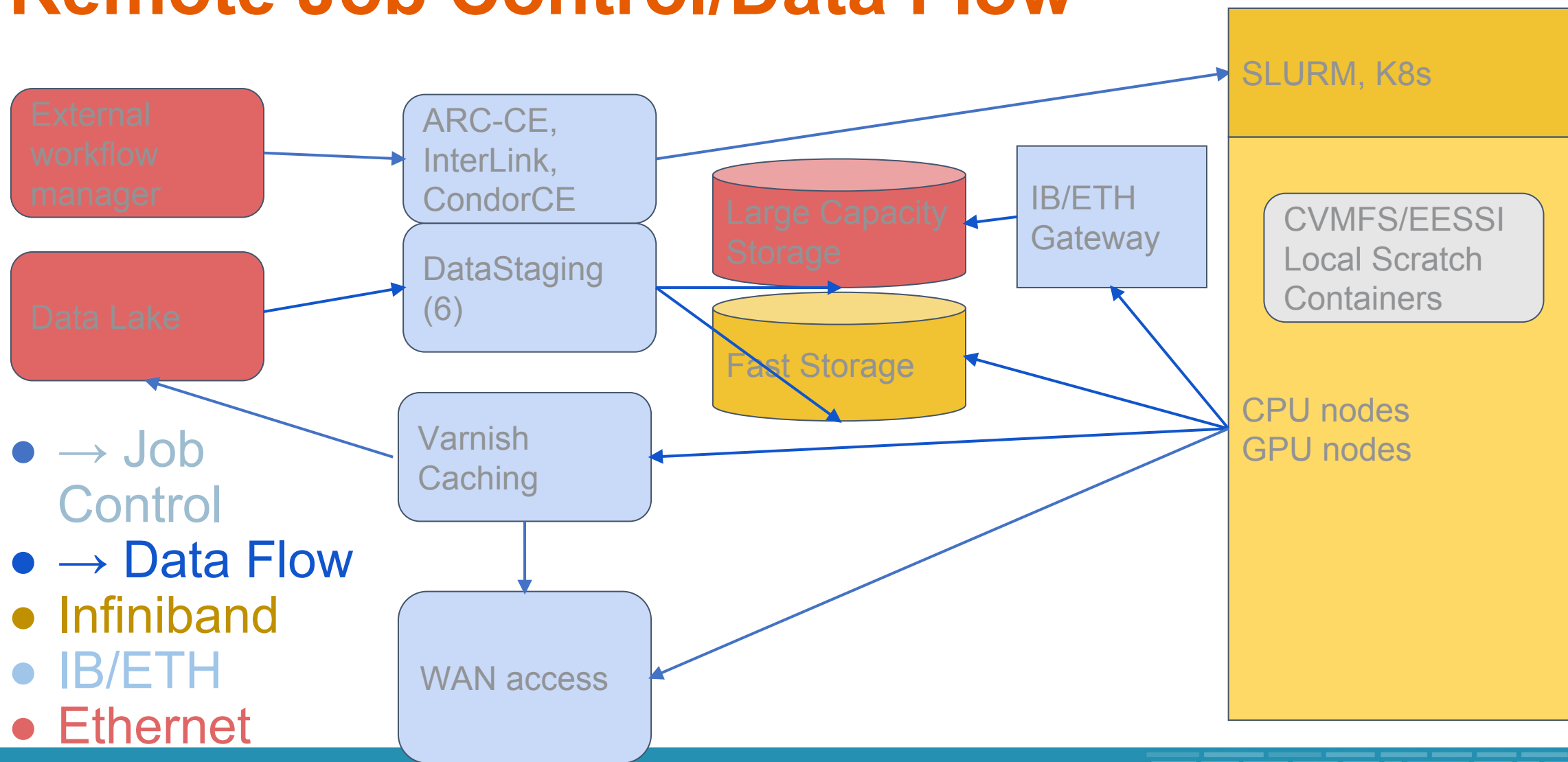


Hybrid architecture

- **HPC/queue SLURM** for efficient utilization
- Cloud interfaces / **K18s** for interoperability
- Automated **configuration** and **provisioning**
- Support for **workflow deployment**
- **Scientific** software and **AI** stack
- (EESSI, CVMFS, registries and repositories, data management)



Remote Job Control/Data Flow



European AI landscape and EuroHPC

- EuroHPC and member countries co-fund the establishment of **AI factories** (with AI-optimized supercomputers) and **AI factory antennas** (providing services but using those same machines):
 - ~ 19 AI Factories so far
 - ~ and 13 antennas where only 9 factories are buying or building a new supercomputer
- Some member states participate and invest in machines in other countries (i.e. Portugal and Turkey in BSC AIF, the LUMI consortium)
- Slovenia will host a new machine, is open to antennas,

And participates in IT4LIA machine (with 5+5 mio



Switzerland's participation is contingent upon the ratification of its accession to Horizon Europe.
 The boundaries and names shown and the designations used on this map do not imply official endorsement or acceptance by the European Union.
 This designation shall not be construed as recognition of a State of Palestine and is without prejudice to the individual positions of the Member States on this issue.
 Administrative boundaries: © EuroGeographics © OpenStreetMap
 Cartography: Eurostat - IMAGE, 05/2025

Why is Slovenia investing in an AI Factory?

Not every EU contry is building and AI Factory.

Ingredients needed: vision, political will, funding support, technical knowledge and community support.

- **A well established and strong AI community** (*since 1970s*)
- **A well established HPC community**

With a national SLING consortium and HPC Vega as the first deployed EuroHPC machine

- **Both communities are collaborating well** (*which is not obvious in Slovenia*)

A wider understanding of infrastructure:

- Not just physical machinery (supercomputing)
- Also data (i.e. Slovenia heads CLARIN, producing linguistic data)

- **AI itself is infrastrucure** (for science, RAZUM as an example competence centre at JSI)

SLAIF – Slovenian AI Factory: projects

Two projects, total worth of 135 mio EUR, co-funded by EC and RS (MVZI)

- AI-optimized supercomputer: 123.32 mio EUR, 7 years since end of 2027
- Co-funded via Digital Europe

Project consortium:



- AI Factory – services and activities, 11.68 mio EUR, 3 year since mid-2025
- Co-funded via Horizon Europe

Project consortium:



SLAIF – target user groups

Industry

- Start-ups
- Small and medium companies
- Large companies
- Public sector
 - Health and medical
 - Public administration
 - Public companies, institutions and agencies
- Research and higher educations (universities, research institutes)

Very different levels of readiness for AI.



SLAIF – applications and priority pillars

Green Transition

Precision farming
Environmental monitoring
Energy optimization
Smart Manufacturing

Health and Biotech

Personalized medicine
Decision support
Medical imaging
Genomics and Diagnostics
Federated learning for analysis of medical data

Digital Society

Language technologies
Media and Creativity
Public Sector
Education and Learning

AI for Science

Support for Research Processes
Life sciences
Material Sciences
Digital Humanities

Summary

- *Modern infrastructure available in Slovenia*
- Considerable resources and effort provided to HEP experiments.
- *HPC Strategy seems to work and is backed by funding*
- *New AI-HPC machine to replace Vega delayed, tendering will start in October*
- *Funding is foreseen and secured for Quantum Computer by 2030 (~10M)*
- *F9 infrastructure is aging, we need to decide how to proceed*

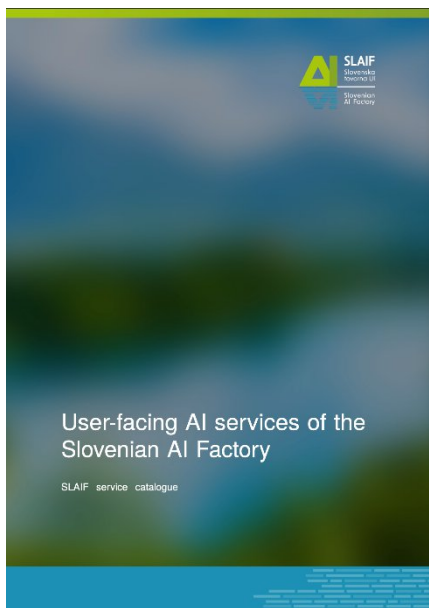
AI for Science

- Research Process Support
 - LLM agents
 - Symbolic regression
 - Automated Hypothesis Generation
- Life Sciences
 - Virtual experiments
 - high-throughput simulations
 - genomics
- Material Sciences
 - New materials for energy
 - Catalysts, Material property prediction
 - Sustainable production
- Digital Humanities
 - LLMs for linguistics
 - History and tools
 - Support for Legal disciplines
 - Research for culture and arts

SLAIF – service catalogue

Main service types:

- User support
- Infrastructure and data
- Development and AI tc
- AI products



Contents

User support	9
Business Processes Review and Automation Using AI	11
Support at Distributed Training of Large-Scale AI Models	12
Support in Model Optimisation: Pruning, Quantisation, and Distillation	14
Support in Proof-of-Concept (PoC) Application Development	16
Technical User Support	18
Infrastructure and data-related services	19
Cloud Platform Service (OpenStack PaaS)	21
Container Image Registry	22
Container Platform Service (Kubernetes)	23
Data Access Services	24
Data Analytics Services	26
Semantic Data Discovery over the SLAIF Catalogue	28
Data Governance Services	30
Data Management / Sharing Services – Connectors for Trusted Sharing	31
Data preparation services	33
Data Processing Services	35
Exploration and Investigation of Disaster Events through the Slovenian EO Data Cube	37
Facebook PE-Video with Slovenian Annotations	39
FAIR-Compliant Metadata Management and Discovery Service	40
Fast Parallel Storage	42
Foundation-Model-Assisted Dataset Bootstrapping from Uncurated Visual Data (Auto-Annotation)	43
HPC Accelerated Compute Service	45
HPC Compute Service	47
Mid- and Long-Term Storage	48
Mid- and Long-Term Remote Access Storage	49
Monitoring Services	50
Orchestration Services	51
Synthetic Data Services	52
Synthetic Genetic Material Data Generation	54
Transformation and Processing Pipeline for Genetic Material Data	55
User Software Repository	57
AI development and deployment	58
Foundation Models for Human Sensing with Support for Fine-Tuning	60
Adaptation of a Large Language Model for Domain-Specific Question Answering	62
Adaptation of a Large Language Model for Domain-Specific Summarization	63
Adaptation of a Large Language Model for a Specific Task using Parameter Efficient Fine-Tuning	64
Adaptation of a Large Language Model for Sentiment and Emotion Classification using Parameter Efficient Fine-Tuning	65
Adaptation of a Large Language Model to a Specific Domain using Instruction Following	66
AI-Based Image Analysis for Industrial Anomaly Detection and Quality Control	67
AI-Based Microscopy Image Analysis	69
Automated Construction of Domain-Specific Knowledge Graphs	71
Change Detection and Description from Remotely Sensed Data	73
Embedding Selection – Intelligent Embedding Selection for Scalable AI Systems	75
ML Pipeline Exploiting Known Attribute Value Uncertainties	77
Model Evaluation for the Local Context Using Linguistic and Cultural Competence Measures	78
Per-Prompt LLM Selection – Automatically Choosing the Best Open Access LLM	80
Preference Optimization Pipeline and LLM-Based Machine Translator for English and Slovene Pairs	82
Satellite and Aerial Image Analysis	83
Temporal, Multi-Modal Foundation Models from Remotely Sensed Information	85
Visual and Interactive AI Workflows (OrangeWeb)	87
Visual and Interactive AI Workflows (OrangeWeb) for Green Transition	88
Visual and Interactive AI Workflows (OrangeWeb) for Health and Biotechnology	89
AI products	90
Access to a Multimodal Language/Sensor Model for Data Analysis	92
Agentic RAG for Administration	94
AI-Assisted Workflow for Grading Scanned Handwritten STEM Exams	96
AI-Based Human Pose Estimation and Kinematic Analysis	98
AI-Based Predictive Analytics for Strategic Resource and Trend Forecasting	100
AI-Based Visual Defect Detection and Quality Assessment for Industrial Production	102
AI-Driven Predictive Maintenance and Performance Optimization for Energy Systems	104
AI-Enhanced Weather Forecasting for Power Grid Operations	106
AI for Health and Biotechnology – AI-Assisted Lab Blood Test Interpretation	108
AI Framework for ECG Analysis	110
AI-Powered Intelligent Document Analysis and Semantic Retrieval	112
AIDA – Agricultural Intelligence and Decision Assistant	114
Automated Classification and Feature Extraction for Experimental Materials Data	116
Crop Growth Monitoring and Analytics Services	118
Domain-Agnostic Object Counting, Detection and Tracking in Images	120
Energy Distribution Optimisation	122
FoodyLLM – AI Copilot for Food Nutrition Intelligence	123
Irradiance Prediction Service for Energy Modelling	125
Multi-Label Classification for Microbial Data	127
Multilingual Idiom Semantic Search and Translation System	128
Observable-Cue Mental Health Risk Prediction & Screening Toolbox	130
Open Model for Slovenian Acoustic Prosody Annotation	132
Pre- and Post-Surgery Primary Brain Tumour Volume Estimation	133

SLAIF – service catalogue



AI development and deployment

Services for adapting, fine-tuning, integrating, scaling, and operationalising AI workflows and models for real use cases.



SLAIF service catalogue



AI development and deployment

Adaptation of a Large Language Model for Domain-Specific Question Answering

Responsible institution
University of Ljubljana

Vertical Cross-domain Availability Q3/2027 Type Pipeline

Target users: Industry, Large Enterprises, Scientific Institutions, Public Institutions

Overview

This pipeline enables the adaptation of a large language model to answer domain-specific questions accurately and efficiently. It is designed to support expert systems, knowledge bases, or FAQ systems in specialized fields. The prerequisites are a sufficient amount of domain-specific data and domain know-how. The pipeline first helps users in preparing data for question answering based on the domain data and provides code and know-how for synthetically expanding the training data based on the provided domain data. Next, it provides the know-how and code for fine-tuning the model with the created instruction-tuning dataset. Finally, it provides code for efficient inference using the adapted model. The pipeline leverages the Slovene LLM GaMS and similar open-source models, tools for generating domain-specific question-answer pairs using NeMo Curator in combination with the GaMS-Instruct approach, LoRA or QLoRA-based supervised fine-tuning using Hugging Face and NeMo, and vLLM for fast domain-specific question answering inference.

Expected outcome

A step-by-step guide for domain-specific question-answering training, along with scripts for data generation, model training, and inference.

Requirements

Requires several NVIDIA GPUs.

Service output

Pipeline for question answering adaptation, including data generation scripts, model training scripts, and inference code.

Specific limitations

Licensing

MIT

SLAIF service catalogue



AI development and deployment

Adaptation of a Large Language Model for Domain-Specific Summarization

Responsible institution
University of Ljubljana

Vertical Cross-domain Availability Q3/2027 Type Pipeline

Target users: Industry, Large Enterprises, Scientific Institutions, Public Institutions

Overview

This pipeline enables the adaptation of a large language model (LLM) to generate high-quality, domain-specific summaries for technical or scientific documents, reports, research papers, or industry-specific content. It is designed to enhance the accuracy and relevance of summaries within a specific domain. The pipeline first helps users in preparing data for summarization (OCR of PDF documents) and converting it into a training dataset. Next, it provides the know-how and code for fine-tuning the model with the created dataset. Finally, it provides code for efficient inference using the adapted model. The pipeline leverages the Slovene LLM GaMS (for Slovene-language summarization) and other open VLMs (e.g., Gemma), open-source OCR libraries such as Marker or small open-source VLMs such as Nanonets, supervised fine-tuning or preference tuning using GRPO with LoRA or QLoRA-based fine-tuning, and vLLM for efficient, low-latency inference of domain-specific summarization tasks.

Expected outcome

A step-by-step guide for domain-specific summarization adaptation, along with scripts for data preparation, model training, and inference.

Requirements

Requires several NVIDIA GPUs.

Service output

Pipeline for summarization adaptation, including data preparation scripts, model training scripts, and inference code.

Specific limitations

Licensing

MIT

62

63

SLAIF – training and skills



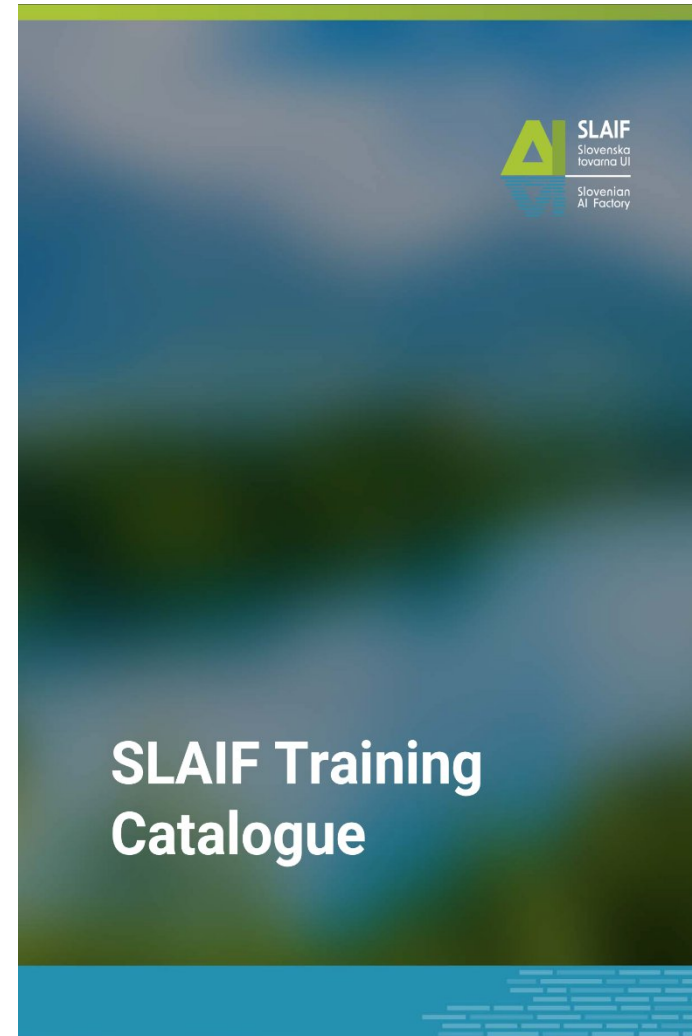
Different training types:

- Workshops
- Webinars
- Training with micro-certificates

Different themes:

- AI methods
- HPC, data and scalability
- Areas of AI use

Different target groups



SLAIF Training Catalogue



Large language models and domain-specific AI for the Slovenian agricultural sector

Course provider:

Jožef Stefan Institute (JSI)

Instructors:

Marko Debeljak (JSI) and colleagues

Content: The event aims to provide agricultural stakeholders with a clear, realistic, and application-oriented understanding of large language models (LLMs) and their potential role in Slovenian agriculture. Participants will be introduced to the rationale for focusing on high-quality, domain-specific data and benchmarks rather than individual model versions, and to the role of the national AI infrastructure (SLAIF) in supporting domain- and language-specific AI solutions. A key objective is to foster informed expectations, trust, and early engagement of future users by openly addressing both the capabilities and limitations of current AI technologies in regulated and practice-oriented agricultural contexts.

The event presents current activities related to the development and application of large language models within the Slovenian agricultural domain in the context of SLAIF. It introduces the fundamentals of LLMs in an accessible manner, with a focus on their applicability, risks, and constraints in agriculture. Participants will be presented with a proposed taxonomy of Slovenian agriculture, including key sectors and subdomains, along with example benchmark tasks designed to evaluate model performance in Slovenian. Preliminary benchmark results will be demonstrated to illustrate how existing models perform on selected agricultural use cases. The event further provides a structured forum for validating assumptions about users, use cases, and sectoral coverage, and introduces a dedicated web application for the collection of high-quality question-answer pairs and other expert inputs to support model adaptation and evaluation.

After the event, participants will understand the importance of domain-specific data, benchmarks, and expert validation for the sustainable development of AI solutions in agriculture. They will have a realistic understanding of what current large language models can and cannot do in the Slovenian agricultural context, and how national AI infrastructure supports their adaptation. Participants will be able to critically assess proposed use cases and benchmarks from a practitioner's perspective and will be equipped to contribute expert knowledge, feedback, and structured data inputs for the continued development and evaluation of AI tools for Slovenian agriculture.

Prerequisites: No technical prerequisites are required. The event is intended for practitioners and experts with domain knowledge in agriculture; prior experience with AI technologies is not required.

Target audience: Agricultural advisers and extension services; Farmers and agricultural practitioners; Agri-food stakeholders and industry representatives; Researchers and developers working on AI for agriculture; Public-sector stakeholders and policy makers

Language: Slovenian; English

Tentative date: Autumn 2026

Venue: ISI Ljubljana/Murska Sobota

Format: Lectures; Hands-on sessions

Duration: 6h

No. of ECTS credits: N/A

SLAIF – training and skills

88

Planned training activities

59

Workshops and practical trainings

16

Webinars

13

Micro-certificates



Safe, responsible and compliant AI for SMEs: From generative AI to the EU AI Act

Course provider:

Jožef Stefan Institute (JSI)

Instructors:

Maja Škrjanc (JSI), Mitja Jermol (IRCAI)

Content: This 2-hour training is designed for SMEs, small and micro companies that are introducing—or considering the introduction of—artificial intelligence, with a particular focus on generative AI. Participants receive a structured overview of core AI concepts (including the distinction between data-driven methods and GenAI, and the role of context, memory and reasoning), alongside a business-oriented discussion of typical risks in practice (e.g., confidentiality and data leakage, GenAI misuse, deepfakes and brand/identity fraud).

FAIR data management for artificial intelligence

Course provider:

Jožef Stefan Institute (JSI)

Instructors:

Panče Panov (JSI)

Learning objectives: 1) Understand AI data assets across the lifecycle: how datasets, labels, dataset splits, features, and evaluation artifacts evolve from collection to reuse; 2) Apply the FAIR principles to AI work: make data and outputs easier to find, access, combine, and reuse (for teams and future projects); 3. Create an actionable DMP for AI projects: a lightweight plan that supports reproducibility, handover, and compliance; and 4) Handle constraints responsibly: recognize sensitive data, ethical considerations, access limitations, and industry vs research expectations.

Efficient LLMops, hosting, and model quantization

Course provider:

Jožef Stefan Institute (JSI)

Instructors:

Boshko Koloski (JSI), Matej Martinc (JSI), Usama Derebashi (JSI), Nikola Marić (JSI), Sašo Džeroski (JSI)

Content: This course covers the operational aspects of large language models (LLMOps), efficient hosting, and model quantization. LLMOps represents an extension of MLOps for generative models; it includes practices for managing the entire lifecycle of LLMs—from selection and customization to deployment and continuous monitoring. Participants learn the differences between MLOps and LLMOps and why generative models require additional procedures, such as robust data pipelines, version control, and flexible resource orchestration.

Large language models and domain-specific AI for the Slovenian agricultural sector

Course provider:

Jožef Stefan Institute (JSI)

Instructors:

Marko Debeljak (JSI) and colleagues

Content: The event aims to provide agricultural stakeholders with a clear, realistic, and application-oriented understanding of large language models (LLMs) and their potential role in Slovenian agriculture. Participants will be introduced to the rationale for focusing on high-quality, domain-specific data and benchmarks rather than individual model versions, and to the role of the national AI infrastructure (SLAIF) in supporting domain- and language-specific AI solutions. A key objective is to foster informed expectations, trust, and early engagement of future users by openly addressing both the capabilities and limitations of current AI technologies in regulated and practice-oriented agricultural contexts.

How to engage with SLAIF?

- **Write to us:** support@slaif.si
 - Meet us: at **events**, lectures, open days
 - ~ → **ticketing & help desk**
 - **What to tell us?**
 - ~ What are you goals and targets → to find common aims
 - ~ What “kind” of client to be → and what we can offer you
 - ~ How deep is your AI journey → what do you need
- consultations → orientation
- ~ **SINGLE POINT OF ACCESS** ← AI CC, NCC for HPC, SLING

Ticket Notification Management

Actions

[Add Notification](#)

[Export Notifications](#)

Filter for Notifications

Just start typing to filter...

Configuration Import

Here you can upload a configuration file to import Ticket Notifications to your system. The file needs to be in .yml format as exported by the Ticket Notification module.

[Durchsuchen...](#) Keine Datei ausgewählt.

☐ Overwrite existing notifications?

[Import Notification configuration](#)

NAME	COMMENT	VALIDITY	CHANGED	CREATED	EXPORT	COPY	DELETE
Ticket create notification		valid	09/21/2020 17:08 (Europe/Berlin)	09/21/2020 17:08 (Europe/Berlin)	+	📄	🗑️
Ticket email delivery failure notification		valid	09/21/2020 17:08 (Europe/Berlin)	09/21/2020 17:08 (Europe/Berlin)	+	📄	🗑️
Ticket escalation notification		valid	09/21/2020 17:08 (Europe/Berlin)	09/21/2020 17:08 (Europe/Berlin)	+	📄	🗑️
Ticket escalation warning notification		valid	09/21/2020 17:08 (Europe/Berlin)	09/21/2020 17:08 (Europe/Berlin)	+	📄	🗑️
Ticket follow-up notification (locked)		valid	09/21/2020 17:08 (Europe/Berlin)	09/21/2020 17:08 (Europe/Berlin)	+	📄	🗑️
Ticket follow-up notification (unlocked)		valid	09/21/2020 17:08 (Europe/Berlin)	09/21/2020 17:08 (Europe/Berlin)	+	📄	🗑️
Ticket lock timeout notification		valid	09/21/2020 17:08 (Europe/Berlin)	09/21/2020 17:08 (Europe/Berlin)	+	📄	🗑️
Ticket new note notification		valid	09/21/2020 17:08 (Europe/Berlin)	09/21/2020 17:08 (Europe/Berlin)	+	📄	🗑️
Ticket owner update notification		valid	09/21/2020 17:08 (Europe/Berlin)	09/21/2020 17:08 (Europe/Berlin)	+	📄	🗑️
Ticket pending reminder notification (locked)		valid	09/21/2020 17:08 (Europe/Berlin)	09/21/2020 17:08 (Europe/Berlin)	+	📄	🗑️