DIGITAL-EUROHPC-JU-2023-AISC-03

MINERVA

European Support Centre for Scalable AI Research and Deployment

Project Number: 101182737

MINERVA Service Portfolio

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.

1 Introduction & Context

The MINERVA project aims to support the development, deployment, and responsible use of high-performance AI across Europe. The project focuses on delivering services that help AI practitioners, in academia, industry and public administrations, make effective use of European HPC resources.

To provide this support in a way that is both structured and adaptable, we distinguish between two key dimensions:

- Service Levels (L1–L5)

  These levels define the thematic areas of support, reflecting the types 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 the format and duration of the support provided — ranging from quick technical help (T1) to long-term engineering collaborations (T4). 📌 *Unlike the levels, these support types are a proposed internal framework to help coordinate activity across partners, allocate resources, and match support formats to user needs.*

This two-dimensional approach enables MINERVA to be both responsive and scalable, helping partners align their contributions with available capacity and expertise, while ensuring high-quality support for the European AI community.

2 Description of the Service Levels

To address the wide spectrum of technical and regulatory challenges faced by the European AI community, MINERVA organizes its service offering into five thematic Service Levels.

L1 – Porting AI Applications & Workflows to HPC

This level supports AI teams in accessing and effectively using HPC infrastructures. Services include assistance with resource applications, adapting applications and workflows to HPC environments, data preprocessing, and storage optimization. Users also receive help with development and execution tools like Jupyter Notebook or Kubernetes, as well as infrastructure-specific documentation.

L2 – Optimization & Scaling of AI Libraries on HPC

Here, the focus is on optimizing the use of AI libraries on HPC systems. This includes scaling workloads for efficient training and deployment, adapting software to hardware accelerators, and addressing regulatory concerns related to responsible AI. Support may also involve adapting large multimodal models and exploring techniques to obtain compact model variants.

L3 – Pre-training of Large-Scale Foundation Models

This level is dedicated to enabling the pre-training of large-scale and open foundation models on HPC infrastructure. Activities include guidance on adapting and scaling models for specific tasks, optimizing performance, integrating diverse data types, and implementing efficient model compression techniques.

L4 – Specialization & Fine-Tuning of Foundation Models

MINERVA helps teams fine-tune and specialize existing models using HPC systems. This includes support for parameter-efficient training, advanced methods like Retrieval-Augmented Generation (RAG), and preference alignment. Efficient evaluation and inference strategies are also part of this level.

L5 – Ethical & Responsible AI Support

Support at this level helps users address challenges stemming from EU regulations and best practices in ethical AI. Services include evaluating datasets for inappropriate content, designing fair training workflows, and advising on AI compliance. This support is delivered in collaboration with external service providers and other EU-funded projects.

3 Description of the Support Types.

To complement the thematic structure of MINERVA services (the Levels), MINERVA proposes to classify service delivery by **Support Type**. This internal framework introduces four types (T1–T4), each corresponding to a distinct mode of response to user needs and partner engagement.

These types define the **format, duration, and depth** of support — from rapid technical troubleshooting to sustained engineering collaboration. The structure helps partners coordinate service delivery, manage resources, and ensure consistent handling of user requests. Each Support Type applies to one or more service levels, depending on the complexity and context of the request.

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 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 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 Evaluation-based selection & Partner availability

3.1 *T1 – Assistance*

Short-term, reactive technical help

Assistance is designed to offer quick, targeted support in response to user issues — typically those related to the use of HPC infrastructure or AI tools. This type of support is triggered through the MINERVA ticketing system.

- Goal: Rapid resolution of specific problems
- Scope: Debugging, cluster adaptation, tool usage, configuration
- Duration: Up to a few days
- Delivery: Remote only
- Trigger: User-submitted ticket

3.2 *T2 – Consulting*

Scheduled advisory sessions on planning and theoretical topics

Consulting provides structured guidance to help users plan their work, understand HPC requirements, navigate ethical and regulatory frameworks, or prepare proposals. It is not hands-on engineering work, but rather expert consultation.

- Goal: Strategic support and clarification on non-code issues
- Scope: Proposal preparation, resource planning, ethical AI, architecture discussions, etc
- Duration: Up to 2 weeks
- Delivery: Remote or optional on-site
- Trigger: User-submitted ticket directly recquiring T2 or follow up of T1/3

3.3 *T3 – Engineering Workshop*

Time-boxed co-working sessions with technical experts

Workshops bring together users and engineers in short, focused sessions to optimize performance, accelerate training workflows, or overcome architectural bottlenecks. These are more involved than consulting, with direct technical collaboration.

- Goal: Accelerate or adapt code and workflows
- Scope: Profiling, parallelization, pipeline optimization, software tuning, etc
- Duration: 1–2 days
- Delivery: Remote or on-site
- Trigger: User-submitted ticket directly requiring workshops or follow up of T1/2

3.4 *T4 – Embedded Support*

Medium- to long-term technical collaboration

This support type involves the integration of a MINERVA engineer into a user project over a longer period. It supports the co-development or scaling of substantial workflows, particularly for advanced AI training or model fine-tuning.

- Goal: Provide sustained technical assistance on ambitious tasks
- Scope: Long training runs, model adaptation, data integration, scaling optimization
- Duration: From a few weeks to several months
- Delivery: Primarily remote, with possible in-person meetups
- Trigger: User-submitted ticket directly requiring T4 or follow up of T1/2/3
    - By request and subject to evaluation and partner availability

📌 *The allocation of T3 and T4 support will depend on human resource availability and internal capacity. Partners are not expected to commit to these formats unless able and willing to do so. T1 and T2, however, are part of the project's baseline service commitment.*

3.5 *Compatibility of Support Types and Service Levels*

Not all support types are equally suited to all service levels. Their compatibility depends on several factors, including the nature of the task, the time required for delivery, and the availability of resources across MINERVA partners.

- T1 (Assistance) and T2 (Consulting) are applicable across all levels, as they are lightweight, flexible formats designed to address most user needs. They are part of the baseline services expected from all partners.
- T3 (Workshops) and T4 (Embedded Support), in contrast, require significant technical involvement and human resources. Their use is best suited to tasks involving in-depth optimization or model development (typically L3 and L4), and they are activated based on partner availability and interest.
- For Ethical & Responsible AI Support, only T1 and T2 are applicable, as these tasks typically involve advice, documentation, or risk assessment, not engineering work.

Service Level\Type	T1 Assistance	T2 Consulting	T3 Workshop	T4 Embedded Support
L1 – Porting & Workflow Adaptation	☑	☑	☑	☑
L2 – AI Library Optimization & Scaling	☑	☑	☑	✗

L3 – Pre-training of Foundation Models	☑	☑	☑	☑
L4 – Specialization & Fine-Tuning	☑	☑	☑	☑
L5 - Ethical & Responsible AI Support	☑	☑	✕	✕

4 Examples of Support Types

To clarify how the different Support Types apply across the Service Levels, the following examples illustrate typical use cases encountered by the MINERVA team. These scenarios highlight both the nature of the user need and the format through which support could be delivered.

The goal is to provide a reference for partners and users, showing how thematic challenges translate into practical engagements.

4.1 *T1 – Assistance*

Example with L1 – Porting

A user encounters an error when submitting a training job on an HPC cluster using Slurm. The issue is resolved via ticket-based support by adjusting their job script and environment setup.

Example with L3 – Specialization

A user faces instability when loading a pre-trained model in their fine-tuning script. A quick investigation helps resolve versioning conflicts in the software stack.

4.2 *T2 – Consulting*

Example with L2 – Scaling

A scientific team has trained an initial model and obtained promising results. They now wish to scale up by increasing dataset size and resolution. A consulting session is held to discuss available HPC resources, best practices in distributed training, considerations for data storage and I/O, and scaling strategies in deep learning (data vs. model vs. pipeline parallelism).

Example with Ethical

A start-up working on medical AI products requests advice on compliance with the EU AI Act. A consulting session walks through potential regulatory risks and documentation requirements.

4.3 *T3 – Workshop*

Example with L2 – Optimization

The workshop focuses on helping users reconfigure their training pipelines using mixed precision and data loading optimizations for better performances and on GPU clusters.

Example with L3 – Fine-Tuning

A team working on LLM specialization participates in a two-day workshop with MINERVA engineers to profile a model using PEFT techniques on multi-GPU setups.

4.4 *T4 – Embedded Support*

Example with L1 – Workflow Adaptation

A research group developing an HPC-based AI workflow is assigned a MINERVA engineer for a few weeks to help refactor their codebase, integrate data preprocessing, and validate performance.

Example with L3 – Pre-training

In collaboration with a national institution, a large-scale effort is launched to pre-train a domain-specific LLM. A MINERVA engineer is embedded in the project for a few months to assist with resource acquisition, workflow development, model scaling, and optimization. Acting as a bridge, the engineer transfers expertise from MINERVA's benchmark data, tools, and HPC best practices into the project team.

5 Request Management and Dispatching

To ensure smooth coordination and equitable distribution of support across partners, MINERVA defines a clear process for handling user requests, tailored to the nature of the support type.

5.1 *Entry Points and Request Routing*

- All initial requests (T1-T4) are submitted to the mail support@minerva4ai.eu, that is connected to a central ticketing system.

5.2 *Allocation of Requests to Partners*

Once received, requests are assessed and routed according to content and context:

- When a request clearly relates to a specific infrastructure, software environment, or partner (e.g., usage of Jean Zay or a national LLM project), it will be forwarded directly to that partner.
- For general or transferable requests, a shared pool mechanism will be used. Partners can volunteer to take requests based on their availability and technical match.

All partners of the MINERVA consortium are expected to contribute to T1 (Assistance) and T2 (Consulting) activities as part of the core WP2 responsibilities. In contrast, the delivery of T3 (Workshops) and T4 (Embedded Support) will be based on each partner's human resource availability and strategic interest. These formats often involve dedicated engineering time and longer planning cycles and will therefore be scheduled and distributed accordingly.

This process allows for dynamic workload balance while respecting partner specialization and preferences.

5.3 *Managing Resource-Intensive Support Types (T3 & T4)*

Support formats like Workshops (T3) and Embedded Support (T4) require significant time and engineering involvement. To avoid overcommitment and ensure quality, these services will be handled on a case-by-case basis, depending on:

- Partner availability at the time of the request,
- Alignment of expertise, and
- Relevance to project priorities or national efforts.

While T1 and T2 support is expected as part of the baseline commitment, T3 and T4 cannot be guaranteed on demand. Requests for these formats will be evaluated collectively, and their scheduling coordinated to match available human resources.

This approach ensures that support remains both high-quality and sustainable across the lifetime of the project.

6 Conclusion

The dual structure presented in this document — Service Levels (L1–L5) and Support Types (T1–T4) — provides a clear and scalable framework for organizing service delivery within MINERVA. While the Levels reflect the technical or regulatory nature of user needs, the Support Types define how partners can respond in practice, based on the depth and duration of engagement required.

This structure enables a consistent, transparent, and efficient approach to coordinating support activities across the consortium. It balances user expectations with partner capabilities and fosters collaboration by providing clear entry points for both simple requests and complex, long-term engagements.

All partners are expected to contribute to the baseline service activities — T1 (Assistance) and T2 (Consulting) — as part of WP2 commitments. More intensive formats, such as T3 (Workshops) and T4 (Embedded Support), are activated based on availability, capacity and interest, ensuring sustainable use of resources while maximizing impact.

By adopting this operational model, MINERVA is positioned to offer high-quality, adaptable, and community-driven support that directly serves the goals of the European AI ecosystem.