

LUMI AI Factory

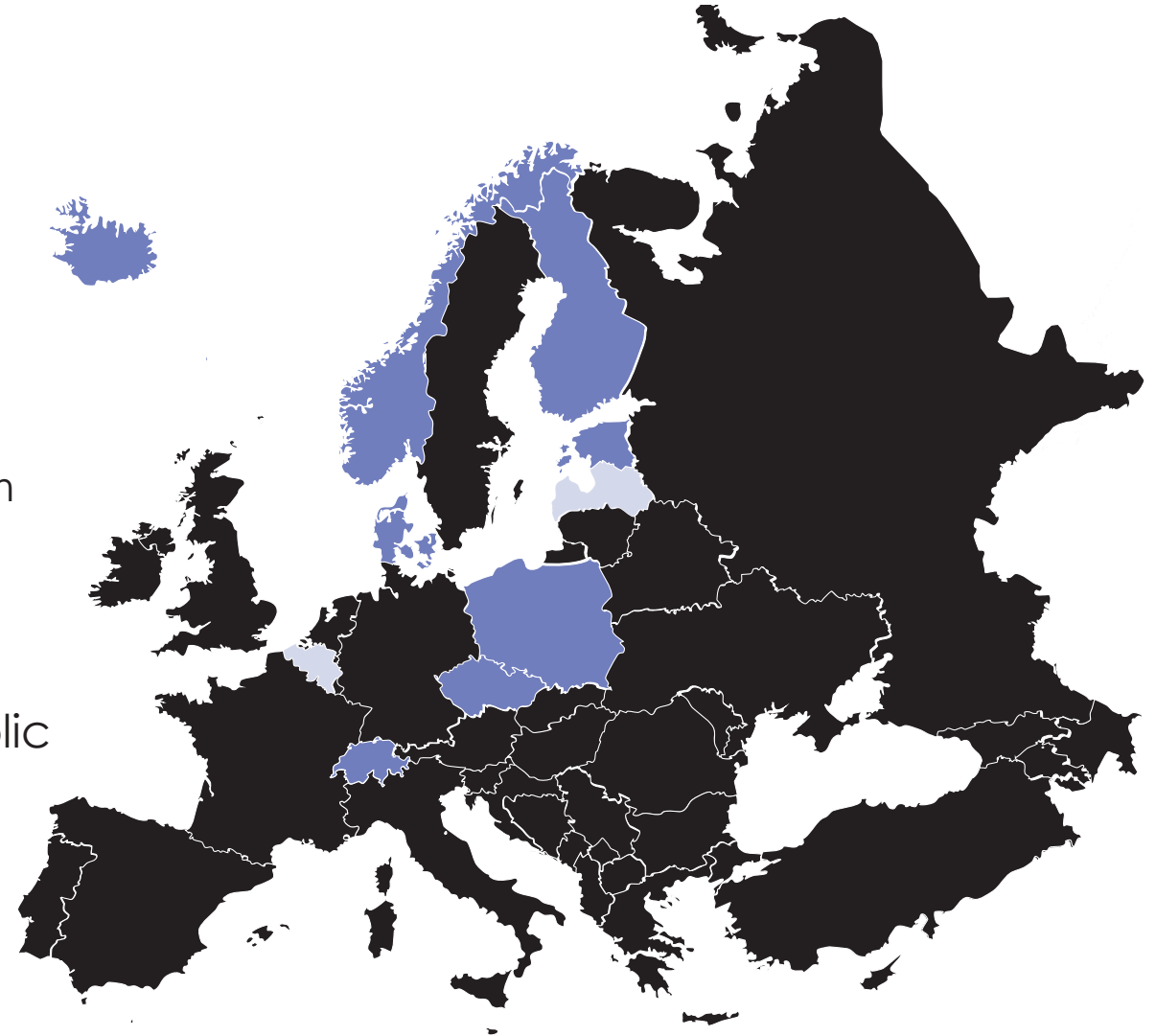
AITTA Inference Platform

LUMI User Coffee 07.10.2026

Lukas Prediger, Marlon Tobaben (CSC, Finland)

LUMI AI factory introduction

- Three pillars
 - **LUMI-AI** – (upcoming) AI-optimised supercomputer
 - AI Factory **Service Center**
 - Industry consultation, other additional services
 - LUMI-IQ – experimental quantum-computing platform
- Funded by the EuroHPC Joint Undertaking and the participating states (650 million euros)
- **Free of charge for startups, SMEs**, research and public sector. Fee-based for large companies.



AITTA

An OpenAI-compatible¹ ML² inference platform powered by LUMI.

Abstracting HPC system complexities - Query state-of-the-art large open language models on LUMI without the struggle (software setup, storage, SLURM and SSH tunnels).

¹ selected endpoints and features only

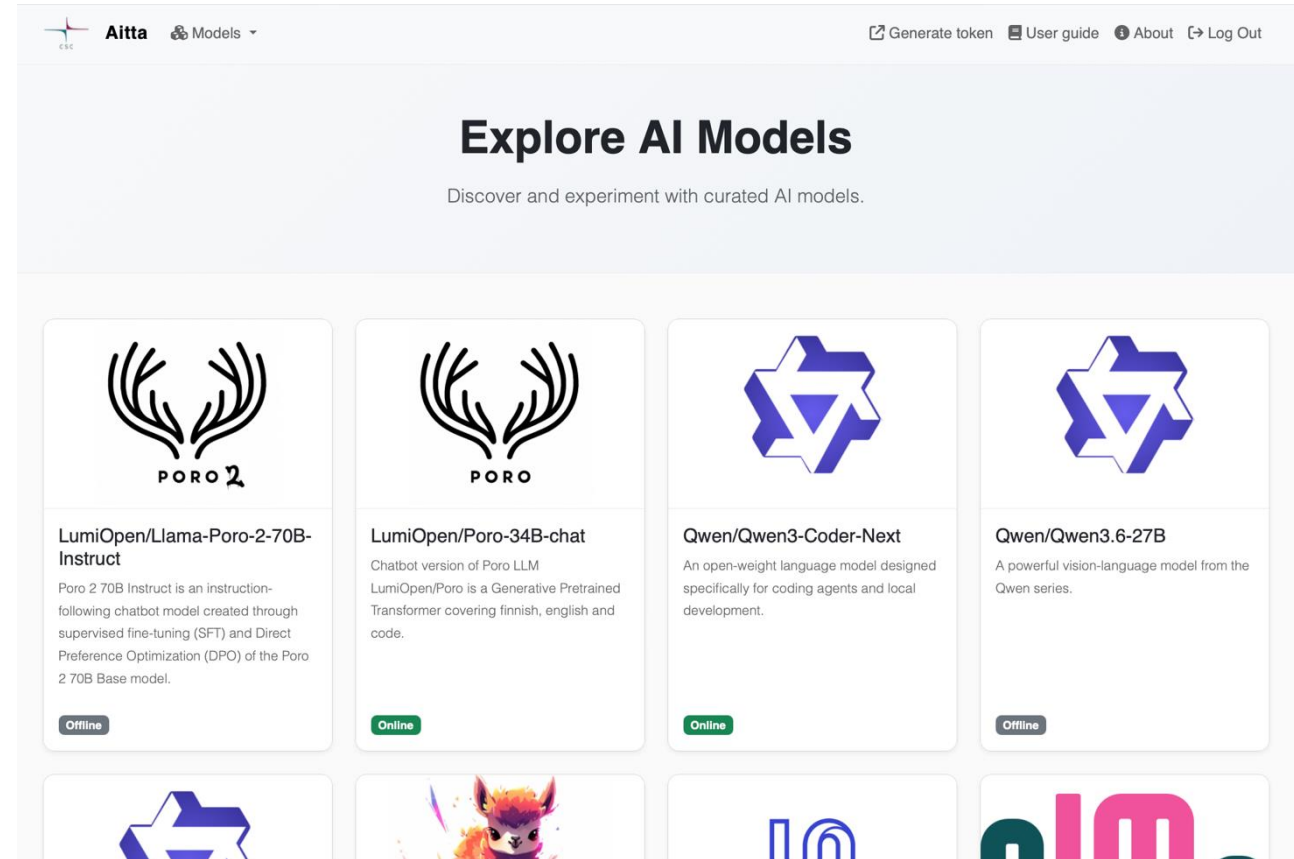
² LLMs

Features

- OpenAI-compatible chat completion and embeddings
 - response streaming
 - tool calls
 - reasoning
- Automatic worker management
 - starting workers
 - load-dependent worker scaling
 - stopping idle workers

Features

- Curated list of open models
 - Google/gemma-4-31b-it
 - openai/gpt-oss-120b
 - Qwen/Qwen3-Coder-Next
 - and more
- Web frontend for chatting
 - zero-hassle interaction with models
- Billed from your LUMI GPU-hours (upcoming)



Where is it?

aitta.csc.fi

Web frontend

docs.lumi-supercomputer.eu/laif/inference/aitta/

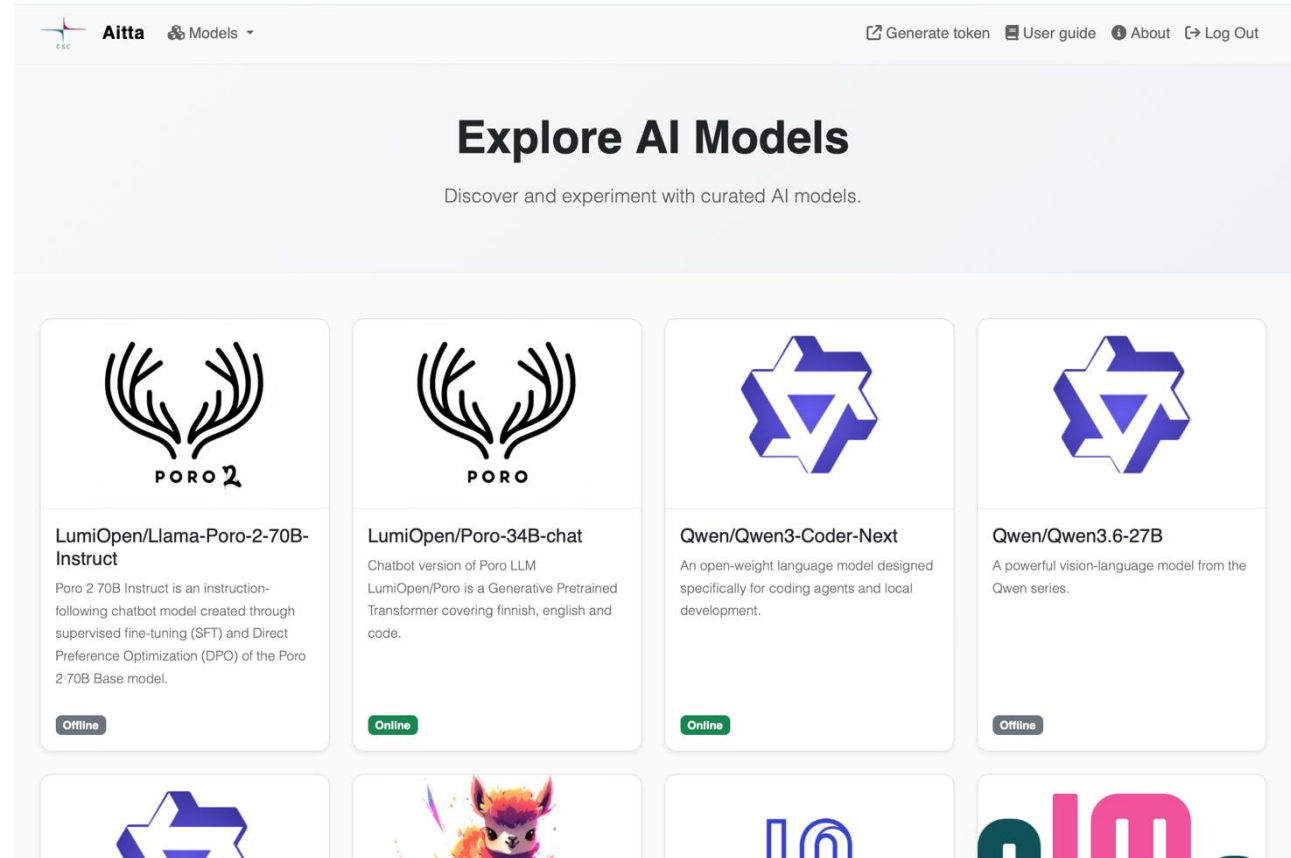
Documentation

aitta-api.csc.fi

API

aitta-auth.csc.fi/myToken

API token generator



What can I do with it?

- Research and explore capabilities of open models
- Explore integrating LLMs in your workflows
- Develop services relying on LLMs
- Offline¹ "production" inference tasks

¹ as opposed to highly-available low-latency interactive

What can't I do with it?

- Expect high-availability
 - Aitta workers subject to SLURM queues and LUMI availability

Responsibility for model usage

The models available via the Aitta inference service are provided "as-is" without any additional safeguards or guarantees for correctness or availability. The user is always responsible for their use of the models and must ensure that their use falls under the [Terms of Use](#) and applicable law.

The screenshot shows the LUMI AI Factory documentation website. The header includes the LUMI logo, the word 'Documentation', and a search bar. A navigation menu lists various sections: Home, First steps, Hardware, Compute, Software, Developing, Storage, Tutorials, Help desk, and LUMI AI Factory (which is underlined). On the left side, there is a sidebar menu with categories: LUMI AI Factory, Services, Data (with a sub-item 'Dataset as a Service'), AI Inference (with sub-items 'Overview' and 'Aitta Inference Service'), Software (with sub-items 'Overview', 'LAIF AI containers', 'Containerized Workflows', and 'Infrastructure for AI agents'), and 'Aitta Inference Service' is highlighted in blue. The main content area features the title 'Aitta Inference Service' and two paragraphs of text. The first paragraph describes Aitta as a machine learning inference service that allows interaction with large open-weight models directly on the LUMI Supercomputer hardware without dealing with technical details. The second paragraph explains that Aitta provides an API for programmatic access and a web frontend with basic chat interfaces. A third paragraph states that Aitta is currently available for all users with access to the LUMI supercomputer. On the right side, there is a table of contents with links to 'Table of contents', 'What can I use it for?', 'Web Frontend - Interactive chat usage', 'Limitations', 'API - Programmatic access' (with sub-links 'Get your access token', 'Inference in Python', 'API reference', 'Rate limits', and 'Capacity limits'), and 'Changelog'.

LUMI Documentation

Home First steps Hardware Compute Software Developing Storage Tutorials Help desk LUMI AI Factory

LUMI AI Factory
Services
Data
Dataset as a Service >
AI Inference
Overview
Aitta Inference Service
Software
Overview
LAIF AI containers
Containerized Workflows
Infrastructure for AI agents

Aitta Inference Service

Aitta is a machine learning inference service that allows you to interact with large open-weight models (i.e., models that are openly available for download e.g. from huggingface.co) directly on the LUMI Supercomputer hardware without dealing with any of the technical details. You just send your inference request to an OpenAI compatible API - Aitta takes care of allocating resources from the supercomputer scheduling system (Slurm), loading the model and forwarding the request to the designated compute node.

Aitta provides an API for programmatic access which allows you to develop or test applications that require inference as a service component. It also enables you to perform automated evaluation of a model you are interested in. Additionally, Aitta provides a web frontend with basic chat interfaces that you can use to interact with the models directly.

Aitta is currently available for all users with access to the LUMI supercomputer. You need a valid LUMI account with access to at least one active LUMI project.

Table of contents
What can I use it for?
Web Frontend - Interactive chat usage
Limitations
API - Programmatic access
Get your access token
Inference in Python
API reference
Rate limits
Capacity limits
Changelog

docs.lumi-supercomputer.eu/laif/inference/aitta/



Aitta

[Generate token](#) [User guide](#) [Log In](#)

Run AI models with Aitta

Aitta provides researchers and developers access to curated AI models running on energy-efficient HPC infrastructure.

[Log in to use Aitta](#)

Run your AI models in LUMI supercomputer with Aitta

AI inference platform Aitta is a general purpose scalable service component of LUMI to host AI models for the research and development purposes.

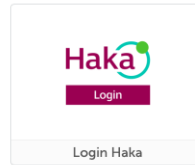
Ease of use

Use LUMI supercomputer powers with easy Web UI and OpenAI-compatible API.

Efficiency

Resource allocation on demand from the European LUMI supercomputer and scalable

Select authentication provider



[Which authentication method should I use?](#)

LUMI AI Factory Login Service

Select Lumi Project

Please select the Lumi project for this session:

☐ 462000131

☐ 462001075

☐ 465002387

☐ 462001302

☐ 462001444

☐ 462001479

☐ 462001488

☐ 462001520

CONTINUE

**Aitta**

 Models ▾

[Generate token](#) [User guide](#) [About](#) [Log Out](#)

Explore AI Models

Discover and experiment with curated AI models.



LumiOpen/Llama-Poro-2-70B-Instruct

Poro 2 70B Instruct is an instruction-following chatbot model created through supervised fine-tuning (SFT) and Direct Preference Optimization (DPO) of the Poro 2 70B Base model.

Online



LumiOpen/Poro-34B-chat

Chatbot version of Poro LLM

LumiOpen/Poro is a Generative Pretrained Transformer covering finnish, english and code.

Online



Qwen/Qwen3-Coder-Next

An open-weight language model designed specifically for coding agents and local development.

Offline



Qwen/Qwen3.6-27B

A powerful vision-language model from the Qwen series.

Offline

google/gemma-4-31b-it

A lightweight, state-of-the-art open model from Google, built from the same research and technology used to create the Gemini models.

Model

About

Chatbot



Hi, can you explain to me what the LUMI supercomputer is?



Reasoning

LUMI is one of the most powerful supercomputers in the world. Located in Kajaani, Finland, it is a massive computational engine designed to solve some of the most complex scientific problems facing humanity.

Here is a detailed breakdown of what makes LUMI special, from its technical power to its environmental impact.

1. What is its purpose?

LUMI was built as part of the **EuroHPC Joint Undertaking**, a project by the European Union to ensure that Europe has the computing power necessary to compete globally in science and technology.

It isn't used for a single task; instead, it is a shared resource. Researchers from across Europe can apply to use its processing power for things like:

- **Climate Modeling:** Creating hyper-accurate simulations of global warming and weather patterns to better predict natural disasters.



Web Frontend Demo – Starting a Model

 **Aitta** 

[Generate token](#) [User guide](#) [About](#) [Log Out](#)

TinyLlama/TinyLlama-1.1B-Chat-v1.0

This is the chat model finetuned on top of TinyLlama/TinyLlama-1.1B-intermediate-step-1431k-3T.

[Model](#) [About](#)

 The model is currently not running.

Start model

Web Frontend Demo – Starting a Model

 **Aitta** 

[Generate token](#) [User guide](#) [About](#) [Log Out](#)

TinyLlama/TinyLlama-1.1B-Chat-v1.0

This is the chat model finetuned on top of TinyLlama/TinyLlama-1.1B-intermediate-step-1431k-3T.

[Model](#) [About](#)

Waiting for computing resources ...

Start model

Web Frontend Demo – Starting a Model

 **Aitta** 

[Generate token](#) [User guide](#) [About](#) [Log Out](#)

TinyLlama/TinyLlama-1.1B-Chat-v1.0

This is the chat model finetuned on top of TinyLlama/TinyLlama-1.1B-intermediate-step-1431k-3T.

[Model](#) [About](#)

 The model is loading. This can take several minutes...

Start model

Web Frontend Demo – Starting a Model

 **Aitta** 

[Generate token](#) [User guide](#) [About](#) [Log Out](#)

TinyLlama/TinyLlama-1.1B-Chat-v1.0

This is the chat model finetuned on top of TinyLlama/TinyLlama-1.1B-intermediate-step-1431k-3T.

 Model

 About

 Chatbot

Getting an API Access Token

 **Aitta** 

[Generate token](#) [User guide](#) [About](#) [Log Out](#)

TinyLlama/TinyLlama-1.1B-Chat-v1.0

This is the chat model finetuned on top of TinyLlama/TinyLlama-1.1B-intermediate-step-1431k-3T.

Model [About](#)

Chatbot

Getting an API Access Token

LUMI AI Factory Login Service

Select Lumi Project

Please select the Lumi project for this session:

☒ 462000131

☐ 462001075

☐ 465002387

☐ 462001302

☐ 462001444

☐ 462001479

☐ 462001488

☐ 462001520

CONTINUE

Aitta Token Bookkeeping

[User Profile Dashboard](#)[Regenerate token](#)

Active tokens can be revoked in the Connected Organisations section of the User Profile Dashboard. Regenerating issues a new token; the old one remains active.

Your Aitta API token

Projects: 462000131

expires at 2027-01-04T06:38:42+00:00 (in 2160h 0m)

```
akjladflkadfslkfdslmak
```



Use this token directly against the API

```
curl -H "Authorization: Bearer akjladflkadfslkfdslmak" "https://aitta-api.csc.fi/model"
```



```
import openai

client = openai.OpenAI(
    api_key="< your token >",
    base_url="https://aitta-api.csc.fi/openai/v1"
)
chat_completion = client.chat.completions.create(
    messages=[{
        "role": "user",
        "content": "Write a novel about Quantum Computing."
    }],
    model="openai/gpt-oss-120b",
    stream=True
)

for chunk in chat_completion:
    if chunk.choices[0].delta.content is not None:
        print(chunk.choices[0].delta.content, end='', flush=True)
```

Managing your Access Tokens


CSC

Personal Data

Connected Organizations

Activity Page

Personal Data

ACTIONS

Attribute	Your value
Account Lock Status ⓘ	false
CSC Username ⓘ	mmustermann
LUMI Projects ⓘ	462000131 462001075 465002387 462001302 462001444 462001479 462001488 462001520

Managing your Access Tokens



Personal Data

Activity Page

Connected Organizations

Find service:

Service

Aitta Token Generator

^

TIMES:

7

LAST ATTRIBUTES

CSC Username
subject
Account Lock Status
LUMI Projects

ACCESS TOKEN

TOKEN EXPIRATION: Mon Jan 04 08:38:42 EET
2027 SCOPE: [openid, csc_user]
ID: _123

REVOKE NOW

Machine Learning Inference Platform

v

Billing

- Usage billed from your project's GPU-hour budget
- Per-model conversion of token usage to GPU-hours
 - Depending on model resources
- Project to be billed selected during login
- Token billing rates will be available via /model and in the UI
- Not yet activated

LUMI AI Factory Login Service

Select Lumi Project

Please select the Lumi project for this session:

☐ 462000131

☐ 462001075

☐ 465002387

☐ 462001302

☐ 462001444

☐ 462001479

☐ 462001488

☐ 462001520

CONTINUE

Authentication

- API requires access token sent via HTTP headers
 - “Authentication: Bearer <your token>”

access token

- for one LUMI project (for future billing)
- 90 day lifetime
- aitta-auth.csc.fi/myToken

How do I use it? (API quick reference)

- /openai/v1/chat/completions
 - OpenAI chat completion request
 - /openai/v1/embeddings
 - OpenAI embedding request
 - /downtimes
 - list known upcoming maintenance breaks
- Use the OpenAI Python client for these
- 

API OpenAPI / Swagger Docs at aitta-api.csc.fi/docs

How do I use it? (API quick reference)

- /model
 - get a list of available models
- /model/<model-url-id>
 - model details and model specific links
- /workers
 - list currently running workers (see which models are active)
- /workers/<model-url-id>/<worker-id>
 - list currently running workers for a model
- /workers/<model-url-id>/<worker-id>
 - detailed status of a worker

Coming Soon / On the Roadmap

- Billing
- Audio Transcription (openai/v1/audio/transcriptions - in testing)
- New models (on an ongoing basis)
- Open Source Release (deploy your own instance for custom models)
- Multimodal generation
- openai/v1/responses endpoint (?)
- Your feature wishes

Support

- For general questions LUMI user support: <https://lumi-supercomputer.eu/user-support/need-help/>
- For AITTA: support@lumi-ai-factory.eu

LUMI

Documentation

[Home](#)
[First steps](#)
[Hardware](#)
[Compute](#)
[Software](#)
[Developing](#)
[Storage](#)
[Tutorials](#)
[Help desk](#)
[LUMI AI Factory](#)

Help desk

[Help desk](#)

[Training and events](#)

[Known issues](#)

[LUMI service status](#)

[Mailing list archive](#)

Help Desk

The LUMI User Support Team (LUST) is here to help!

If you have any questions or problems using LUMI, please do not hesitate to contact us. You may find our contact forms for different issues by clicking the 'Contact us' button:

Contact us

The LUMI helpdesk is open every weekday from Monday to Friday except during [Finnish bank holidays](#):

	CE(S)T	EE(S)T	UTC (Summer)
Start	8:00 AM	9:00 AM	7:00 AM (6:00 AM)
End	6:00 PM	7:00 PM	5:00 PM (4:00 PM)