Towards Federated GenAI Model Serving and Knowledge Retrieval in EOSC-ARENA

Estíbaliz Parcero[1] · Miguel Caballer[1] · Germán Moltó[1] · Judith Sáinz-Pardo[2] · Ignacio Heredia[2]
Pablo Fernández[3] · Manuel Otero[3] · Fotis Psomopoulos[4] · Konstantinos Votis[4] · Nikolaos Laloumis[4]

[1] Instituto de Instrumentación para Imagen Molecular, Universitat Politècnica de València
[2] Instituto de Física de Cantabria (IFCA), CSIC–Universidad de Cantabria
[3] SCORE Lab, I3US Institute, Universidad de Sevilla
[4] Centre for Research & Technology Hellas (CERTH)

We aim to make generative AI usable as a research service by connecting flexible model deployment, scientific knowledge and measurable service quality.

IBERGRID 2026 | Part of the AI4EOSC initiative

Funded by the European Union

Motivation: Researchers need more than access to a model

A research team wants to analyse a collection of publications with a generative model.

01



Choose a model service

Public services offer model access. Some workflows need a specific model, repeatable configuration, or control over data processing.

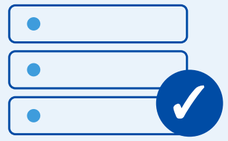
02



Connect the evidence

Useful answers need relevant sources. Retrieve material from publications, repositories and the team's own collections, then supply it to the model.

03



Depend on the service

Researchers need reliable availability and response times. Operators need performance evidence and clear service objectives.

The challenge: usable models, relevant scientific evidence and dependable services.

EOSC-ARENA: bringing GenAI into open science

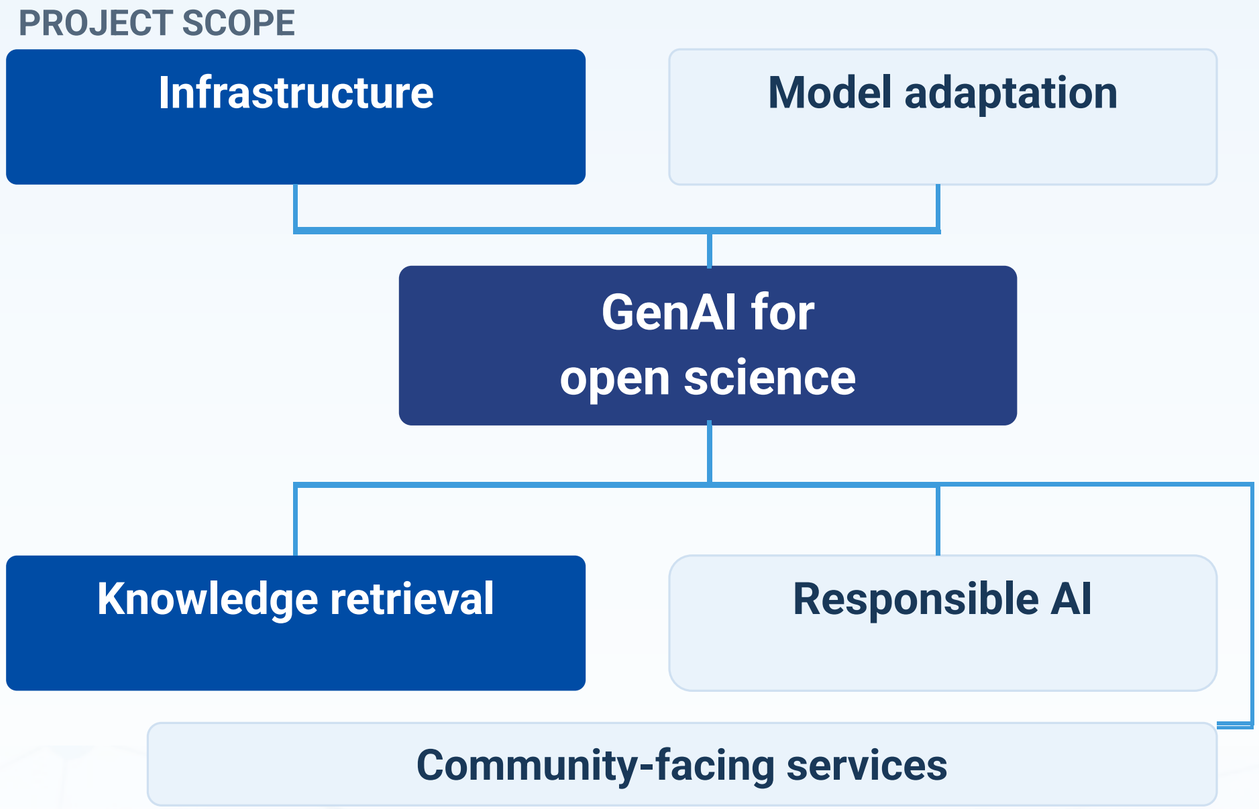
The project connects a broad scientific ambition with the services needed to deliver it.

A scientific assistant

EOSC-ARENA aims to build a sovereign, generative and agentic AI environment within EOSC, supporting research from literature review to analysis, reporting and provenance.

Built with research communities

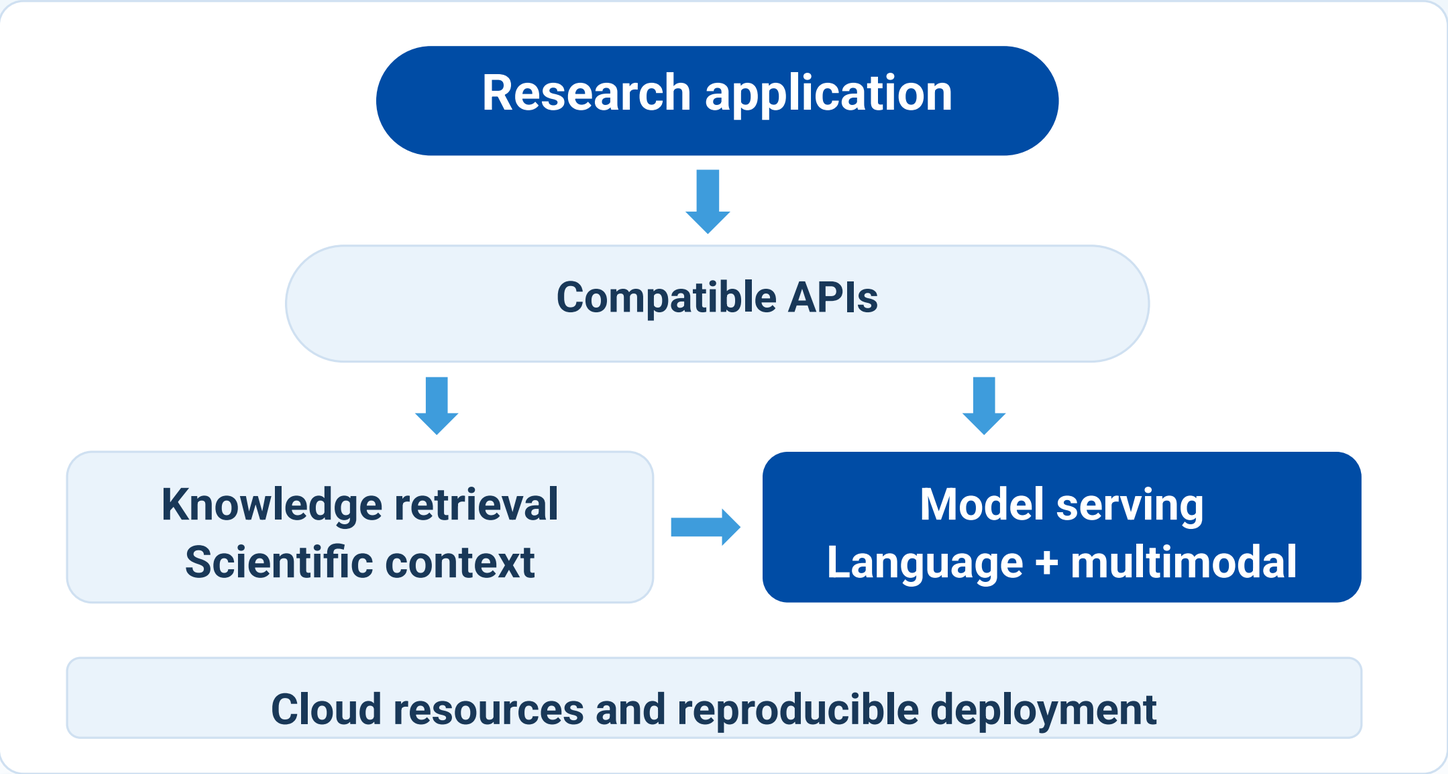
The project builds on AI4EOSC. Twelve use cases will help co-design and assess its services, alongside work on responsible AI and interoperability.



Our focus: model serving and knowledge retrieval, supported by measurable service quality.

How the technical layer fits together

Conceptual integration: connect the research workflow, model access and scientific information.



Connect the workflow

Compatible APIs connect research applications to model and retrieval services.

Prepare the environment

Infrastructure Manager provisions resources and configures the required services.

Observe and manage

Metrics and SLA policies guide service-quality decisions.

The contribution brings serving, retrieval and service management into one technical foundation.

Build on OSCAR and LLM4EOSC

Existing AI4EOSC developments provide complementary starting points for research services.

OSCAR · UPV

A Virtual Research Environment for scientific workflows with event-driven execution and on-demand containers.

GENAI SERVING ROADMAP

OSCAR → **KServe** → **vLLM**

KServe manages model serving on Kubernetes;, a runtime like vLLM executes inference. The roadmap includes Predictive Inference Services and Generative Inference Services. http://oscar.grycap.net

LLM4EOSC · IFCA

Managed access to open models through the AI4EOSC service. Researchers can use chat or integrate model calls into their own applications.

Chat / API → **vLLM backend**

An OpenAI-compatible API supports application integration. Embeddings enable document retrieval.

https://ai4eosc.eu/products/#llm

Build on existing capabilities and connect them through a documented integration trajectory.

Match the deployment to the research workload

Three deployment modes balance demand, operational support and control.

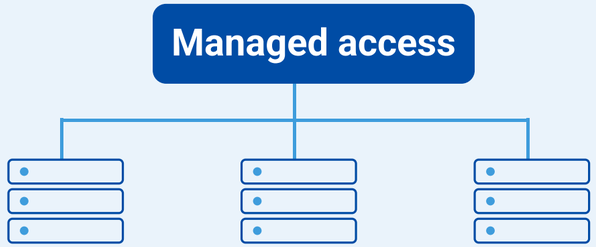
Serverless inference



EVENT → EXECUTION → RESULT

Serverless inference runs processing in response to events or requests.

Managed federation



PARTICIPATING ENVIRONMENTS

Managed instances span participating environments. Teams get compatible services and shared operational support without running each deployment.

Self-hosted private



A dedicated environment gives teams control over models, configuration and sources, along with responsibility for operating the service.

Choose the deployment mode around the workload, data needs and service objectives.

Make deployments repeatable across environments

Infrastructure Manager turns reusable descriptions into provisioned and configured services.

Deployment recipe

TOSCA specifies infrastructure requirements. Ansible roles define service configuration.

Infrastructure Manager

IM provisions resources and configures services on the selected computing backend.

https://www.grycap.upv.es/im

Repeat the setup across scientific, public or private cloud resources.

Reusable environments for research

The same approach supports model-serving environments and retrieval systems prepared for a team's own information sources.

Future direction:
EOSC EU Node scenarios

Reusable deployment recipes make service setup easier to repeat, share and adapt.

Connect model responses to scientific evidence

Illustrative workflow: ask which methods are described in a collection of publications.

Prepare sources: extract, split and index content.

Research question

Sources	Retrieval	Context	Served model	Answer
Publications and repositories	Select relevant passages	Question + retrieved material	Generate a response	Evidence that can be inspected

RAG supplies scientific context

RAG supplies retrieved passages as model context so researchers can inspect the evidence behind a response.

MCP connects tools and resources

MCP gives applications and agents a standard way to access retrieval tools and information sources.

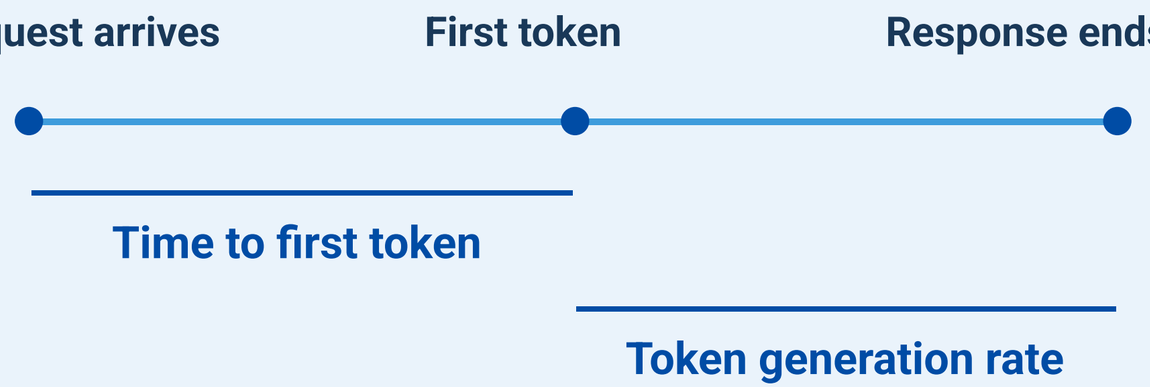
Project direction: managed retrieval and environments for users' own sources.

Retrieval connects the model workflow to scientific sources, including the user's own information.

Make service behaviour visible and measurable

Illustrative signals: connect the researcher's experience with model and infrastructure behaviour.

What happens during one request?



Interactive work cares about when output starts; batch work also cares about sustained throughput.

Service signals

Availability, request success and response time describe the application experience.

Resource signals

Operational and performance metrics help us understand service behaviour and guide management decisions.

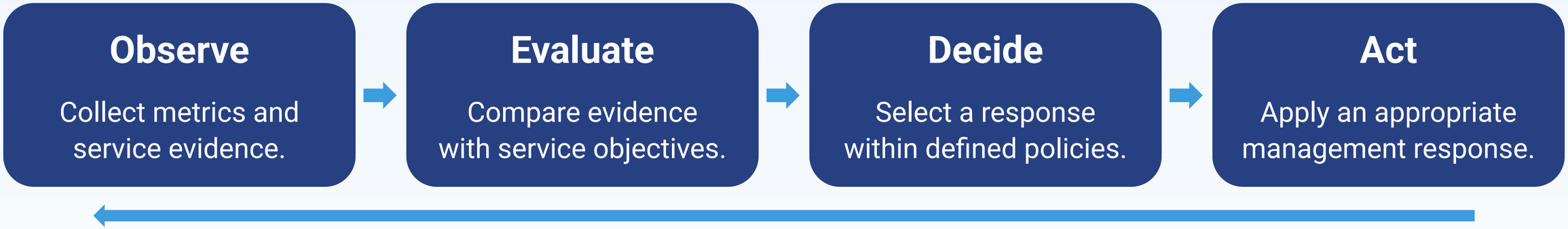
Workload objectives

Interpret signals against the service expectations of each workflow.

Observability supplies the evidence; service objectives tell us how to interpret it.

Use service objectives to guide management actions

Proposed agentic SLA loop: evaluate operational evidence and select responses within explicit policies.



Governify

Collects evidence and evaluates agreement objectives to make compliance inspectable.

SLA Wizard

Translates SLA descriptions into API gateway settings, including usage rate limits.

Example: detect capacity pressure, evaluate the policy, act if authorised, then reassess.

Agentic service management is an integration objective built around evidence, policies and feedback.

Demonstrate the integration in research workflows

Proposed evaluation: move from reusable components to a repeatable and measurable service.

Deploy → **Connect** → **Exercise** → **Evaluate**

Deploy	Connect	Exercise	Evaluate
Repeat the setup	Call compatible APIs	Run a research workflow	Inspect quality and behaviour

Serving and interoperability

Can we repeat the setup and reproduce API behaviour across representative backends?

Evidence and retrieval

Does retrieval find relevant material, and can researchers inspect the sources?

Service quality

Does the service meet workload objectives, and what changes after a policy-guided action?

The evidence must come from the integrated workflow, with explicit evaluation criteria.

Conclusion

Our contribution connects model access, scientific evidence and measurable service quality.

Usable models

Flexible serving and repeatable deployment make language and multimodal models accessible across clouds.

Scientific knowledge

Retrieval brings scientific evidence into model workflows, using shared resources or a team's own sources.

Manageable service quality

Observability and service-level objectives support informed operational decisions and policy-guided actions.

A foundation for future agentic research workflows in EOSC-ARENA

Together, these components provide a foundation for connected GenAI services in research

Acknowledgements

EOSC-ARENA

AI Research Enhancement through Networked Agents

This work has been supported by the EOSC-ARENA project, funded by the European Union under Horizon Europe Grant Agreement No. 101292597 (HORIZON-INFRA-2025-01).

Thank you