# Call for Pilot Foundations: Helmholtz Foundation Model Initiative

# 

# < Project Acronym: Project title >

< Participating center 1, center 2, center 3, external partner 1, 2,…>

Submission deadline 16.09.2024 9:00 am CEST

Project Description

* 1. Abstract

Max. 250 words.

* 1. Key personnel

List of names and contact info of PIs, their host institutions, and principal expertise (e.g. AI, HPC, research domain). Please denominate lead PI(s) for principal contact.

* 1. Project Objectives

Describe the principal project objectives, success criteria, and how the intended Foundation Model will catalyse future research.

* 1. Approach and expected outcome

Please describe the specific approach you will take to meet each of the above objectives. Make reference to:

* What data will be used and how its extent and complexity is expected to give rise to a Foundation Model.
* How the proposed consortium merges domain, AI, and compute expertise to build domain-leading Foundation Models, including proof of ability to scale to large-scale HPC.
* Please describe the expected outcome of your approach, as well as its expected scientific and technical innovation potential and impact in the field of AI research.

1. Technical requirements
   1. Data sources and management

Provide details about data provenance, access, and availability.

* 1. Model Architectures and Algorithms, Code

Provide a technical description of your Foundation Model approach, i.e., model architecture, training & tuning paradigms, etc.. Describe if and how existing state-or-the-art approaches will be employed or improved, and whether new approaches will need to be developed within the project.

* 1. Compute resource requirements

Describe which HPC system you will be using in each phase of model development, and which architecture (CPU, GPU, mixed,…).

If applicable, list any non-standard prerequisites (e.g., any special or proprietary codes/packages/libraries, special hardware, etc.) and how they will be provided on the chosen HPC system.

Provide quantitative estimates of compute- and storage resources in terms of the quantities requested (CPU/GPU hours, storage amounts etc). A proof-of-concept that your code(s) perform at scale on the available HPC infrastructure(s), e.g. via scaling plots, memory- and I/O requirements may be included. The requirements of the Gauss Centre for Supercomputing can serve as a guide ([Sec. 5.1 and Sec. 5.2](https://jards.gauss-centre.eu/PublicFiles/Description/GCS-Template-project-application.pdf)).

A reliable, and convincing plan to obtain the required compute resources must be provided. This plan is crucial in the evaluation of the proposal. If possible, refer to previous successes in obtaining large-scale HPC resources.

Details may be provided in appendix 2 (outside the page limit), but must be referred to and comprehensively summarised here.

1. Project planning, implementation, and monitoring
   1. Work plan with Milestones, Deliverables, Dependencies

Please describe work packages with milestones and deliverables that measure and document project progress, including their due dates. All items must correspond to items listed in the workpackage/deliverable tables and Gantt chart to be provided in the appendix.

We recommend to plan work packages to align with three principal project phases, with principal deliverable targets preferably at 12, 24, and 36 months after project start. The first two years could be devoted to building the fully functional foundation model, measuring success on downstream tasks, and making the model available to the scientific community, whereas the third year could focus on assessing model behavior in depth and extending its scope in a high-risk-high-gain fashion.

Expected deliverables:

* A trained foundation model publicly available in a trusted repository.

Recommended deliverables include but are not limited to:

* Models adapted to several downstream tasks, each publicly available in a trusted repository.
* Quantitative evaluation and benchmarking against baselines for at least three downstream tasks, publicly available with a DOI.
  1. Risk analysis

Provide a risk analysis for the project, focusing on dependencies between milestones, deliverables, workpackages, and discuss mitigation of risks arising from failing tasks and work packages.

* 1. Compute Budget monitoring

Please describe how you plan to monitor and document the use of the required compute budget.

* 1. Data Monitoring

Please describe how you plan to monitor data use in the project; In particular, please address the question how you will prevent data leakage (e.g., leaking training data in model testing).

1. Broader Impacts

Please reflect how the broader impacts of the successful project could affect the field, stakeholders, and society at large.

Discuss the benefit that the Foundation Model will bring to the domain, i.e., how it could advance research and enable addressing research questions that are difficult to access at the current state of the art.

Discuss who potential stakeholders are, in research and outside, and how you will approach them for dissemination of the project results to maximize the impact of your research.

Discuss potential societal impact and benefit of the project, also considering expected long-term impact.

1. Sharing of Project Results
   1. Helmholtz Open Science Policy

All data, models, code should be shared according to FAIR principles. Shared results should fulfil the obligations of the [Helmholtz Open Science Policy](https://os.helmholtz.de/en/open-science-in-helmholtz/open-science-policy/): “as open as possible and as closed as necessary”. Maximum openness is a principal focus of the HFMI.

* 1. Code

Describe how the code developed in this project will be shared to maximize utility for the scientific community.

* 1. Models

Describe how the model(s) resulting from this project will be shared. Describe plans to make the trained models publicly accessible, whether they will be deposited in a trusted repository, under which license.

5.3 Data

Describe how the training data for the model will be published. If applicable, describe any restrictions to public release of the training data.

5.4 Roll-out to the community

Describe plans to roll out the models to the scientific community for productive use, including consideration on modalities of access and documentation for research users.

6. Budget

6.1 Requested FTEs

Please provide a detailed plan for the use of the requested funding, including job description and hosting center for each requested FTE.

Note the requirement to submit a letter by the CEO/board of directors of each participating center underwriting funding of the project in funding period 2026/27.

7. Collaboration Plan and Team Science

Please describe how you will ensure efficient and effective team-work and communication in your consortium. Provide a concept how interaction of members will be supported, in particular in the light of the interdisciplinary and likely distributed nature of the project. For example, describe a plan for regular virtual and/or physical meetings on various levels (e.g. PIs only, or whole consortium, or subgroups; at regular intervals and/or ahead of project inflection points like e.g. milestone or deliverable due dates, roll-out events).

Define project speaker and coordinator roles.

Describe a credit attribution/authorship policy for research products like publications, code, models, talks, etc., considering the different standards across disciplines.