HELMHOLTZ

Research for grand challenges.

Call for Pilot Foundations: Helmholtz Foundation Model Initiative – Proposal preparation Instructions –

Main Proposal

See the **proposal template** for the mandatory structure and content of the proposal. Please retain the headings in the proposal template. Subheadings may be introduced as needed, but their use should be kept to a minimum. The bibliography of references is to be added in the appendix, not in the main proposal.

Maximum length is for the proposal is **10 pages.** Appendices listed below do not count into that page limit but have their own page limits.

Proposal language is English. All content, main proposal and appendices must be formatted in DIN A4 page, margins at least 1.27 cm on all sides, with consecutive page numbers in the main proposal starting on the first content page.

Content is to be written in Arial font, 11pt minimum, single line spacing minimum, and single line spacing minimum.

The use of figures and tables is encouraged where useful. Figures should be numbered and referenced in the text. Captions should be 8pt minimum.

Page limits, page dimensions, font sizes minima etc. are mandatory. Failure to adhere to those minima may result in delayed processing or rejection without review.

Mandatory appendices

All appendices are mandatory unless stated otherwise.

1. Cited references

Bibliography of all references used in the proposal.

2. Compute infrastructure overview

Documentation of availability and access to adequate compute infrastructure.

3. List of Work Packages and deliverables

a. (one page) Tabular view of work packages indicating the leading center as well as a descriptive title, work package budget and a timeline, referencing the proposal section 3.1.

b. (one page) Tabular view of the planned deliverables that lists the associated work package, the responsible project partner and the delivery month. Indicate which deliverables will be produced with or used by external partners.

c. (one page) Gantt chart for with work packages, tasks, deliverables, milestones, dependencies etc.

4. Financial Plan

a. A breakdown of the allocated budget for each work package defining their aggregate in-kind and cash contributions for the entire funding period.

b. A breakdown of the annual budget for each participating center according to type of funding (staff and expenses).

5. CVs of Participating PIs and potential candidates for intended positions

CVs should be provided that clearly reflect a proven record of expertise in the research domain, in AI/ML, and HPC for key personnel. Provide:

a. CVs of the principal investigators (2 pages each max.). List up to ten research products that are most relevant to the project. Research products include not just peer-reviewed papers but could also describe software packages, datasets, policy papers, contribution to standards, patents, and could include accomplishments like entrepreneurship and sustained industry/business collaborations. Provide evidence for the relevance and impact of research products (which may go beyond citation metrics).

b. If available, CVs of prospective candidates for a specific position (2 pages each max.).

6. Confirmation of default commission

A declaration, usually a signed letter, by the CEO/board of directors of each participating center must be included, which guarantees that they are willing and able to support the project in case the central funds for 2026/27 are not available for that period.

Proposals failing to provide this letter will be rejected without review.

7. Ethics and compliance assessment

An assessment on potential ethical concerns or negative societal impact. Please consult with your Institutional Review Board(s) and report the outcome.

If the project involves sensitive data, describe the process of data handling and compliance.

In general, it is recommended to assess the proposal according to the NeurIPS ethics guidelines and briefly report on that.

8. Signed LoS from external partner providing resources (if applicable)

If your consortium includes external partners, provide a signed Letter of Support that lists the resources that are to be provided by the partner and confirms that the partner commits to providing them, in case the project will be selected for funding by the committee.

Online Submission

The main proposal and all mandatory appendices should be submitted as a **single PDF** via email to <inkubator@helmholtz.de>.

The following information is to be entered in the online submission tool:

a. Abstract (must be identical with the abstract in the proposal)

b. Up to 10 keywords

c. Names and contact information of the lead center's (coordinator) and further centers' Principal Investigators
d. Names of participating centers
e. Budget

Further inquiries

For further inquiries, please contact via email:

- Information and Data Science team, inkubator@helmholtz.de
- Florian Grötsch florian.groetsch@helmholtz.de

Appendix: Questionnaires

These questionnaires aggregated key information pertinent to establishing competitive consortia. They are primarily intended to facilitate matchmaking between domain experts and AI experts who would like to participate in the HFMI but have not yet identified suitable consortial partners.

The questionnaires could also be useful for established consortia in sharpening their profile and developing the key points of their proposal.

The domain questionnaire is intended for domain scientists looking for AI and HPC partners. The AI questionnaire aims at AI experts looking for domain experts and HPC partners.

Domain Questionnaire

1. Participating PIs

Please list the names and center affiliations of all PIs who contribute domain expertise, data and/or downstream tasks. Please indicate the prospective coordinating PIs, their prospective Speaker, and the prospective hosting center(s) of the requested Domain team. Please list your (international) collaborations with partners from science and industry (as part of this consortium or prospective ones).

2. Scientific Background

2.1 Scientific Background

Please provide a summary of your domain, the data at hand and what it captures, associated research questions, and the envisioned impact of the foundation project in the context of related work and the current state-of-the-art.

2.2 Preliminary Work

Please describe previous/preliminary work, especially significant contributions in the field of data analysis and AI, of all contributing PIs as relevant for the proposed project, including previous approaches for analyzing the data at hand.

3. Data Description

3.1 Data type:

Please describe the data type(s) and dimensionalities (e.g., 2-dimensional RGB images; 2d hyperspectral satellite images with 10 channels, integer-valued vectors of length 10.000, …)

3.2 Data sources:

Please list the groups in Helmholtz as well as external collaborators and/or consortia who will contribute data. Please indicate whether you have the appropriate licenses and rights for the data.

3.3 Data acquisition:

Please describe, per data source from 3.2, where and by whom the data were acquired. Please describe the provenance documentation status of the data. Please describe the modalities/ instruments with which the data is/was acquired. Please describe the acquisition process. In

particular, please describe to what extent instruments can be parametrized, and if acquisition-specific parameters are documented as metadata.

3.4 Data set size:

Please specify the amount of data available to date:

- the (average) size of an acquisition (e.g., for an image, the number of pixels and bits per pixel)
- the number of acquisitions at hand
- an estimate of the information content per acquisition (e.g., relevant signal in ca 10% of the data)
- the number of independent samples, as well as dependent acquisitions therein (e.g., 1000 distinct materials samples, with 10 acquisitions per material sample at different temperatures)

and use these values to provide aggregate values. In case of distinct parameter settings for an individual instrument type, please describe how diverse these settings are in the data at hand. If warranted, please specify the above values per (standard) parameter setting of an instrument.

3.5 Domain-specific Metadata:

What metadata is associated with your data? Is there domain-specific metadata? If so, please describe the content and format. Does the metadata follow a common ontology used in your domain?

3.6 Data Annotations:

Is the data, or some of it, (expert-)annotated? If so, what has been annotated, how was it annotated, and who performed the annotations? What tools or processes were used for data annotation? Is there provenance documentation and/or metadata available about the annotation process?

3.7 Data examples

Please provide snapshot visualizations of data (or, should the data be hard to visualize, please provide exemplary raw data) for all modalities (or at least for a subset of modalities that is representative of the diversity of data in the set). Please describe the content of the data. Likewise, please provide associated metadata as well as (visualizations of) associated annotations if these exist.

3.8 Data set volatility:

How frequently is new data added to your data set or existing data updated? / Do you anticipate significant growth or changes in the data in the near future?

Data Curation

3.9 Data Quality / Data Curation Status:

How was the data quality assured during collection? Has the data been cleaned or curated? Describe all necessary data quality assurance processes and resp. state of completion. Is there documentation on completed or on-going cleaning/curation processes? Are there known quality issues in the data that still need to be addressed? If so, please describe issues

as well as how resp. curation will be documented. Are there known biases or confounders in the data? If so, how are they documented?

3.10 Required Personnel for Data Collection and Curation:

How many FTEs would be necessary to collect and curate all data, i.e. to make it "AI-ready"? (estimate) Note that a prototype of the data set for initial trainings should be available at most 6 months after the project starts.

Data Access

3.11 Accessibility

Where is the data located? (Please describe per data source from 3.2 if warranted.) Is the data deposited in a trusted repository? If so, does it have an identifier? Is there metadata associated that facilitates data discovery? If so, please describe.

Are there plans to deposit the data, or is it already deposited, in an existing NFDI or any open access repository? Please specify which one, and if possible name a respective contact person.

In which format is the data stored? With which storage mechanism(s)? Who can access the data and how / what is the access protocol?

3.12 Licensing / usage restrictions / ethics issues

Is the usage of the data (or parts of it) restricted? If so, please describe how and why.

Do you have, or do you need to acquire, the consent of third parties to be able to provide and enable exploration of the data?

Under which license is / will your data be made available (for model training only, or also publicly)?

Does the license allow the data to be transferred to another location (e.g., off-site HPC)?

Is there an embargo applied towards public availability, e.g. to give time to publish? If so, please specify why and how long this will apply.

Are there, or do you foresee, any ethics- or legal concerns with data sharing? All results should in accordance with the Helmholtz Open Science Policy as open as possible and as closed as necessary. Please check with the Ethics guidelines of your domain or the NeurIPS code of ethics guidelines.

4. Tasks

4.1 Downstream tasks

Describe several highly impactful, concrete downstream tasks that you expect can be solved by a foundation model trained on your data set.

4.2 Output type

Please specify the type and dimensionality of the kinds of output data you seek to extract from your data, individually for each of the above-mentioned downstream tasks.

4.3 Evaluation metrics:

What are the evaluation metrics for measuring progress/success in each downstream task based on the above-described types of output?

4.4 Evaluation data:

Are there “ground truth” annotations available to evaluate the above-described task-specific metrics?

4.5 Output examples

Please provide visualizations / examples of outputs per downstream tasks. Examples can stem from existing evaluation data, or if this does not exist, be manually sketched or exemplified for data samples provided in 3.6

5. Community

5.1 Helmholtz community:

Please list groups in Helmholtz not listed as contributing PIs who acquire or work with data that you would deem similar to yours, as well as groups who pursue similar research questions.

5.2 Broader community:

Please list further groups or institutions (Helmholtz and external) who may benefit from /use the envisioned foundation model.

5.3 Diversity of data in the broader community:

Please describe the diversity of relevant instrumentation and resp. data acquired in your global community; Are there other, distinct instruments used in your community to acquire similar data? Please describe as far as possible, including an assessment how widespread the resp. instrument is and how distinct the resp. data is to yours.

5.4 External collaborators:

Are there external collaborations or partnerships you have established or envision in the context of this initiative? Please explain the benefit of including external partners to the application.

AI Questionnaire

1. Participating PIs

Please list the names and center affiliations of all PIs who contribute AI expertise. (In case of overlap with Use Case Quest., Sec. 1, please duplicate here). Please indicate the prospective coordinating PIs, their prospective Speaker, and the prospective hosting center (yes, just one! hosting of the AI team ≠ ownership of the project!).

2. Scientific Background

2.1 Scientific Background

Please describe the scientific background of the proposed project, including related work and the current state-of-the-art.

2.2 Preliminary Work

Please describe previous / preliminary / significant work of all contributing PIs as relevant for the proposed project, including previous experience with the data type at hand, the application domain at hand, as well as previous experience with large-scale HPC in the form required for the project.