HELMHOLTZ
Open Science

HELMHOLTZ
Metadata Collaboration

Forum

HELMHOLTZ
Research Data Commons

Enhancing Research Data Workflows for and with AI

Report

HELMHOLTZ
Research Data Commons

Imprint

The online version of this publication can be found at:
https://doi.org/10.5281/zenodo.17265958

Authors

Constanze Curdt, Steffi Genderjahn, Marc Lange, Christine Lemster, Mathijs Vleugel, Elisa Jones, Stefan Kesselheim, Dmitriy Kostunin, Alexandre Strube

Edited by

Marc Lange, Christine Lemster

Published by

Helmholtz Open Science Office and Helmholtz Metadata Collaboration (HMC)

Contact

Helmholtz Open Science Office
c/o GFZ Helmholtz Centre for Geosciences
Telegrafenberg, 14473 Potsdam
E-Mail: open-science@helmholtz.de

Status

November 26, 2025 | Version 1.0

Licence

HELMHOLTZ
Research Data Commons

Contents

Abstract

The Forum Helmholtz Research Data Commons serves as a collaborative format to improve research data management at the Helmholtz Association. This report documents the forum on Enhancing Research Data Workflows for and with AI, held on 30 September 2025. It featured presentations on data readiness for artificial intelligence (AI) applications from a practitioner's perspective, on the Helmholtz' large language model (LLM) service BLABLADOR and on a use case of LLM applications in astronomy. A pitch on possible applications of AI in data management commenced a subsequent discussion. The discussion highlighted both the opportunities and challenges of enhancing research data workflows through AI, and showed that the issue demands not just technology, but a holistic approach focused on improving data quality and sustainability.

HELMHOLTZ
Research Data Commons

Background

The Helmholtz Metadata Collaboration (HMC) and the Helmholtz Open Science Office hosted the second iteration of the Forum Helmholtz Research Data Commons on 30 September 2025, this time focusing on: Enhancing Research Data Workflows for and with AI. In the online event, colleagues from the AI team at the Jülich Supercomputing Center provided practical insights into preparing research data for AI and showcased their AI tool BLABLADOR, a free and privacy-aware Helmholtz AI LLM service, and its possible applications in research data analysis. Following on, a research data use case from astronomy developed by researchers of the Deutsche Elektronen-Synchrotron DESY, which employs BLABLADOR's capabilities, was presented. The event included an open discussion round on the topic with the speakers and the approx. 75 participants, which was commenced with a pitch on possible applications of AI in data management by colleagues of the GEOMAR Helmholtz Centre for Ocean Research Kiel. The subsequent discussion highlighted both the opportunities and challenges of enhancing research data workflows through AI, and showed that the issue demands not just technology, but a holistic approach focused on improving data quality and sustainability.

Research Data Commons is a joint recurring forum for the exchange and discussion of research data-relevant topics at Helmholtz, initiated in 2024 by the Helmholtz Metadata Collaboration (HMC) and the Helmholtz Open Science Office. The events are open to employees from all Helmholtz centers to share their experiences and approaches around specific focus topics.

HELMHOLTZ
Research Data Commons

Program

Time	Program	Speaker
10:00 - 10:10	Welcome and Introduction	Mathijs Vleugel | Helmholtz Open Science Office, Head Constanze Curdt | Helmholtz Metadata Collaboration
10:10 - 11:00	Talks and Presentations	moderation: Marc Lange | Helmholtz Open Science Office
10:10	Is Your Data Ready for AI? A Practitioner's Perspective	Stefan Kesselheim | Forschungszentrum Jülich (AI Team @ Jülich Supercomputing Center)
10:25	BLABLADOR - The experimental Helmholtz AI LLM server	Alexandre Strube | Forschungszentrum Jülich (AI Team @ Jülich Supercomputing Center)
10:35	Use Case: LLM Applications in Ground-Based Gamma Astronomy	Elisa Jones & Dmitriy Kostunin | Deutsche Elektronen-Synchrotron DESY
10:45	Q & A	
11:00 - 11:50	Discussion	all participants moderation: Constanze Curdt | Helmholtz Metadata Collaboration
11:00	Pitch: Possible Applications of AI in Data Management	Carsten Schirnick & Gregor Börner | GEOMAR Helmholtz Centre for Ocean Research Kiel
11:05	Open Discussion: Enhancing Research Data Workflows for and with AI	
11:50 - 12:00	Closing Remarks	Constanze Curdt | Helmholtz Metadata Collaboration

Talks and Presentations

Is Your Data Ready for AI? A Practitioner's Perspective

Stefan Kesselheim | Forschungszentrum Jülich (AI Team @ Jülich Supercomputing Center)

The recent successes of AI and Machine Learning have been possible only as a consequence of published datasets. Benchmark datasets such as ImageNet[1] have been developed as tools to measure the progress in the field, and have become the quasi-standard. Recently, extremely large data collections, such as The Pile[2] and Fineweb2[3] are the fertile ground for the development of open source LLMs. Scientific datasets such as the Protein Data Bank are the basis for breakthroughs such as AlphaFold[4]. Highly visible, high-quality datasets in the right context can contribute to significantly advance a scientific field. In my talk, I'll discuss the success factors for datasets and different strategies to make them visible for AI method experts. Breakthroughs in AI are steered by the challenges that the scientists are trying to solve. Help them find the most exciting ones.

BLABLADOR – The Experimental Helmholtz AI LLM Server

Alexandre Strube | Forschungszentrum Jülich (AI Team @ Jülich Supercomputing Center)

BLABLADOR is the experimental LLM server for Helmholtz. [5] It is an open-source inference server optimized for serving predictions from customized scientific LLMs. It was developed and is operated by Helmholtz AI Jülich to make scientific LLMs more accessible to Helmholtz scientists. BLABLADOR hosts both open-source models and models created within Helmholtz, and is accessible via a web-based chat interface and an API. BLABLADOR enables a collaborative ecosystem for scientific LLMs, in which researchers can add their pretrained LLMs to a central hub, which thereby become available for querying by other researchers.

---

[1] Deng et al. (2009). ImageNet: A large-scale hierarchical image database. 2009 IEEE Conference on Computer Vision and Pattern Recognition. https://doi.org/10.1109/CVPR.2009.5206848

[2] Gao et al. (2021). The pile: An 800gb dataset of diverse text for language modeling. arXiv. https://doi.org/10.48550/arXiv.2101.00027

[3] Penedo et al. (2025). Fineweb2: One pipeline to scale them all -- adapting pre-training data processing to every language. arXiv. https://doi.org/10.48550/arXiv.2506.20920

[4] Jumper et al. (2021). Highly accurate protein structure prediction with AlphaFold. Nature, 596. https://doi.org/10.1038/s41586-021-03819-2

[5] BLABLADOR is available here (for Helmholtz-internal use): https://helmholtz-blablador.fz-juelich.de

Use Case: LLM Applications in Ground-Based Gamma Astronomy

Elisa Jones and Dmitriy Kostunin | Deutsche Elektronen-Synchrotron DESY

We present a multi-agent application for the next-generation Cherenkov Telescope Array Observatory, designed to automate the generation of Pydantic Python models directly from free-text descriptions or structured files.[6] It also explores the use of the multi-agent framework AutoGen, as well as minimal function tools available in new OpenAI interfaces, incorporating a feedback loop to verify and refine generated code before user presentation, streamlining the workflow for astrophysical data management. The app is baked with Blablador's GPT-OSS, as the best performing model for the task out of selected models.

Discussion

Pitch: Possible Applications of AI in Data Management

Carsten Schirnick and Gregor Börner | GEOMAR Helmholtz Centre for Ocean Research Kiel

Currently, queries to large databases still require specialist knowledge of the database and its query language. AI applications can support this process by simplifying queries and making data more accessible. This pitch will outline the ongoing developments in the context of the data from the German research vessels.

Open Discussion: Enhancing Research Data Workflows for and with AI

The discussion centered on enhancing research data workflows through AI, highlighting both opportunities and challenges. A key theme was the need to bridge the gap between technical data language and natural language, potentially improving accessibility for researchers and data understanding. Concerns were raised about a potential loss of competencies if users become overly reliant on AI without understanding the underlying logic and important syntax (for refining database queries using SQL, for example). Beyond simply understanding data,

[6] See also:
Kostunin et al. (2025). Enhancing the development of Cherenkov Telescope Array control software with Large Language Models. arXiv. https://doi.org/10.48550/arXiv.2510.01299
Kostunin et al. (2025). Agent-based code generation for the Gammapy framework. arXiv. https://doi.org/10.48550/arXiv.2509.26110

participants explored applications of AI to improve data itself – refining legacy data, enhancing metadata quality, and automating improvements like assigning persistent identifiers and ontologies – addressing a significant hurdle in data discovery and usability. Challenges mentioned for successful AI implementation were insufficient resources and lacking investment in necessary infrastructure. Collaboration with industry was suggested as a potential solution, though acknowledged as complex given the diverse needs within the research community. Existing initiatives, like Helmholtz AI and those within the Helmholtz Metadata Collaboration were positively noted, with the caveat of the difficulties of scaling these to address the unique requirements of different research areas. Ultimately, making data 'AI-ready' requires not just technological solutions, but a concerted effort to improve data quality, accessibility, and understanding, alongside sustained investment and tailored collaborative approaches.

Appendix: Presentation Slides

Is Your Data Ready for AI? A Practitioner's Perspective

by Stefan Kesselheim | Forschungszentrum Jülich (AI Team @ Jülich Supercomputing Center)

BLABLADOR – The Experimental Helmholtz AI LLM Server

by Alexandre Strube | Forschungszentrum Jülich (AI Team @ Jülich Supercomputing Center)

The slides can be found on Alexandre Strube's personal page: https://strube1.pages.jsc.fz-juelich.de/2025-09-30-talk-helmholtz-rdc

Use Case: LLM Applications in Ground-Based Gamma Astronomy

by Elisa Jones and Dmitriy Kostunin | Deutsche Elektronen-Synchrotron DESY

HELMHOLTZ
Open Science

HELMHOLTZ
Metadata Collaboration

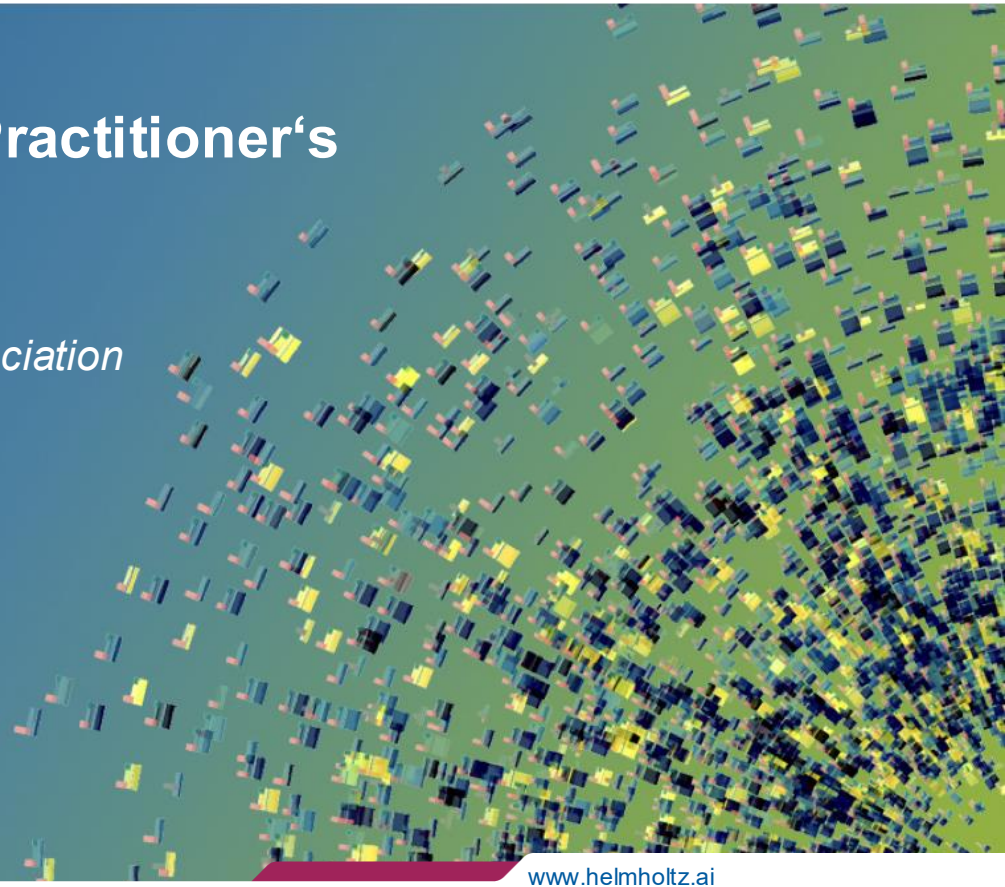
HELMHOLTZ
Research Data Commons

HELMHOLTZ AI | ARTIFICIAL INTELLIGENCE COOPERATION UNIT

Is Your Data Ready for AI? A Practitioner's Perspective

Stefan Kesselheim

Forschungszentrum Jülich – Helmholtz Association

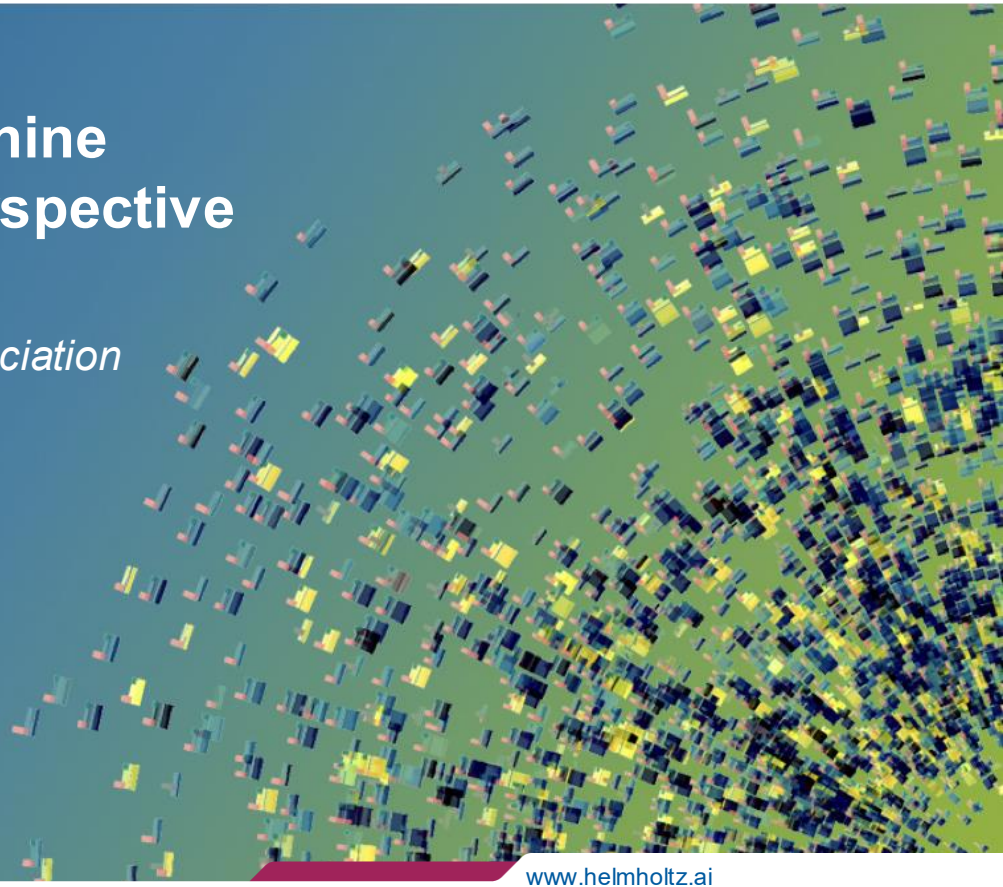


HELMHOLTZ AI | ARTIFICIAL INTELLIGENCE COOPERATION UNIT

Is Your Data Ready for ~~Ai~~ Machine Learning? A Practitioner‘s Perspective

Stefan Kesselheim

Forschungszentrum Jülich – Helmholtz Association

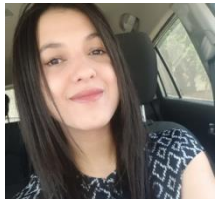


Simulation and Data Lab Applied Machine Learning

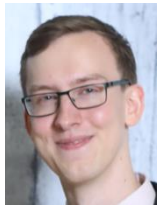
The team


Stefan Kesselheim


Alina Bazarova


Sabrina Benassou


Anoop Chandran

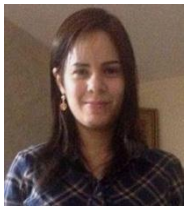

Jan Ebert


Alexander Strube


Fritz Niesel


Oleg Filatov


Emile de Bruyn


Rania Briq


Jiangtao Wang

Data	Collect, select, transform, process
Question	Formulate, refine, transform
Method	Select, develop, modify, improve
Implementation	Create, develop, optimize, parallelize
Communication	Write, speak, present, reflect

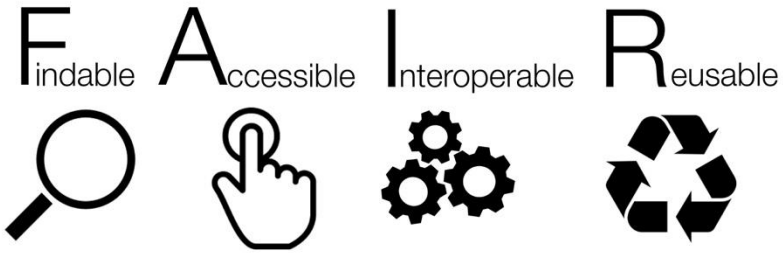
Is your data ready for AI?

Two possible goals

1. You would like to get started with AI methods yourself.
2. You want to draw attention your data and have others develop AI methods for them.

Is your data ready for AI?

A special flavor of FAIR



Addressing sufficiently general questions?

Python access? Other data sources?

How can non-experts understand the question?

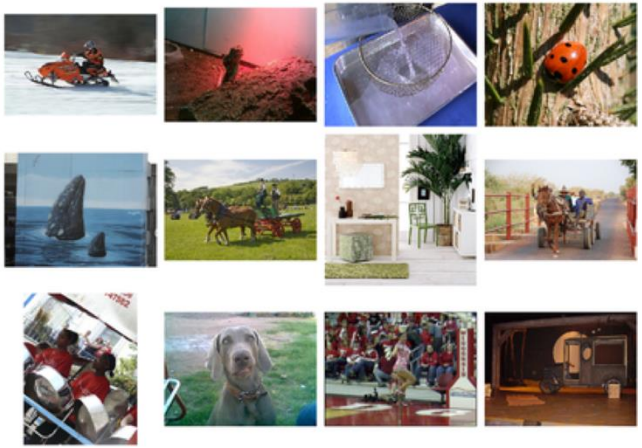
Is the data visible?

ImageNet Large Scale Visual Recognition Competition

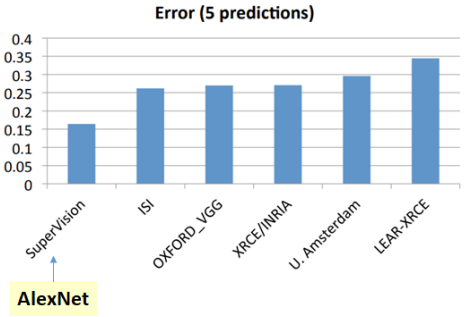
AlexNet (2012)

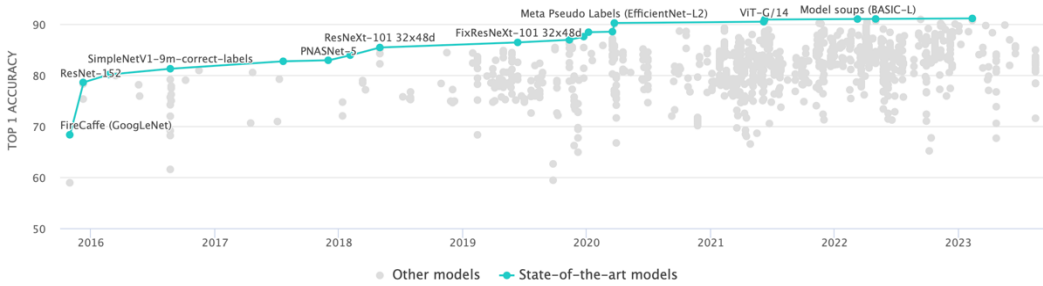
ILSVRC (ImageNet Large Scale Visual Recognition Competition)

- 1.2 Mio Images
- 1000 Classes



Ranking of the best results from each team





https://medium.com/coinmonks/paper-review-of-alexnet-caffenet-winner-in-ilsvrc-2012-image-classification-b93598314160

Is Your Data Ready for AI?

NeurIPS Dataset and Benchmarks Track

- Part of the main NeurIPS conference
- Submission is single-blind, and the review process is open during the discussion phase
- “We welcome submissions that detail advanced practices in data collection and curation that are of general interest”
- Citations six random papers of 2021: 16, 32, 1, 13, 140, 19

NEURAL INFORMATION PROCESSING SYSTEMS

https://neuripsconf.medium.com/announcing-the-neurips-2021-datasets-and-benchmarks-track-644e27c1e66c

Is Your Data Ready for AI?

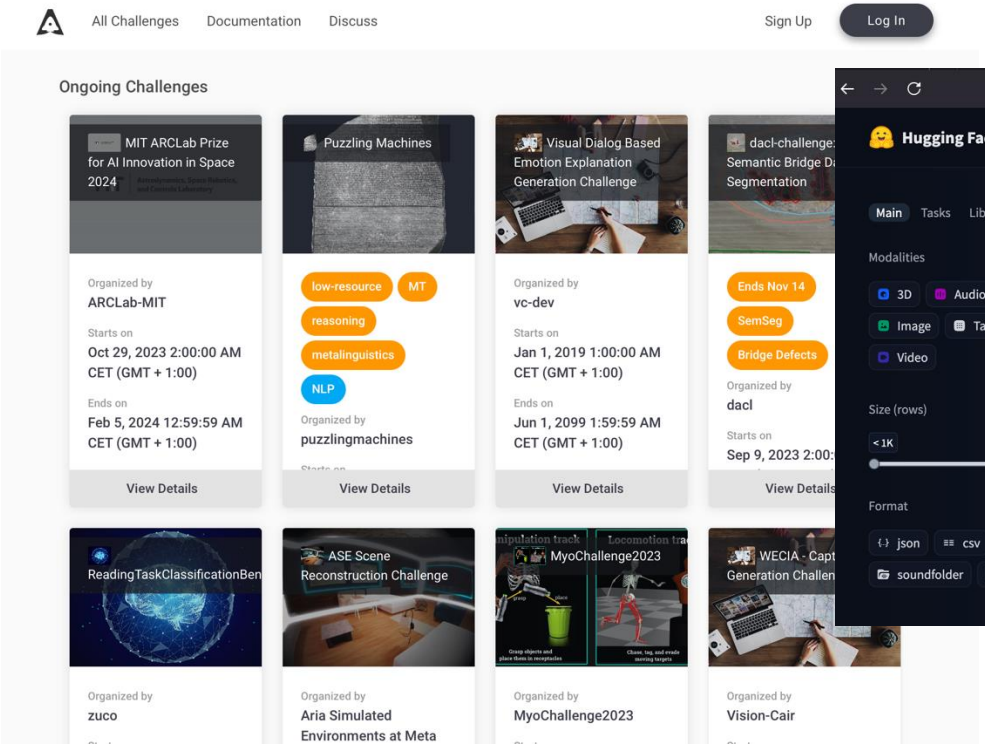
NeurIPS Competition Track

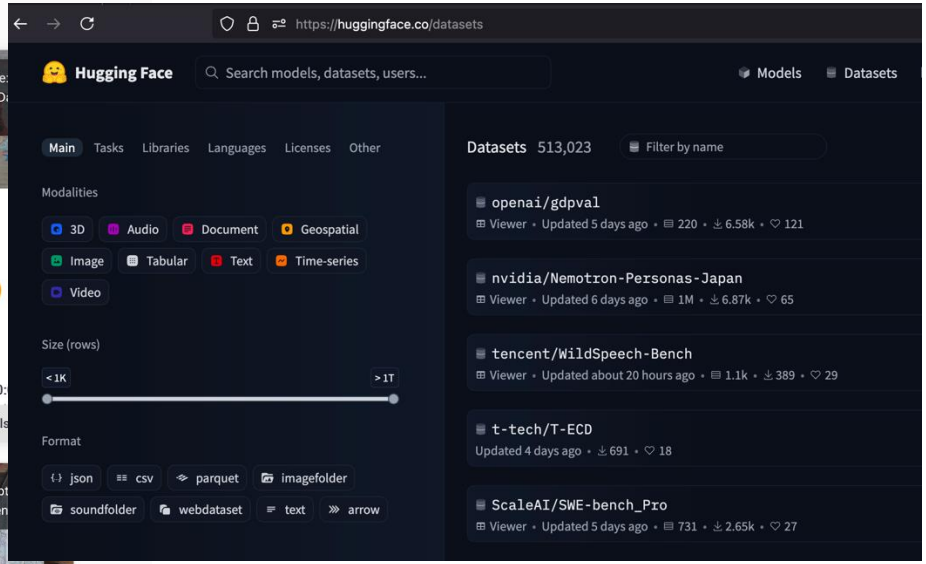
- “Competitions should be motivated by clear scientific questions and curiosity. Impact, originality, and relevance to the NeurIPS community will all be considered. Tasks that include humanitarian and/or positive societal impact are highly encouraged, although other topics relevant to the NeurIPS community are also welcomed.“
- “Feasibility of the task chosen, baseline availability, sufficient data for training and testing algorithms to solve the proposed task, soundness of the evaluation criteria, and clarity and fairness of the competition rules will all be evaluated.”
- “Competition organizers should propose a timeline for running the competition to ensure participants have enough time to contribute high-quality entries.”

NEURAL INFORMATION PROCESSING SYSTEMS

Is Your Data Ready for AI?

Eval.ai & Huggingface





https://huggingface.co/datasets

Yadav, Deshraj, et al. "Evalai: Towards better evaluation systems for ai agents." *arXiv preprint arXiv:1902.03570* (2019).

Is Your Data Ready for AI?

Python readiness

galaxy-datasets Public

main 4 Branches 17 Tags Go to file

mwalmsley Merge pull request #45 from mwalmsley/dev

.github/workflows	very weird, not updating pa
data	add galaxy_mnist as debug
derived_data	swap gzcd for gzh2o
galaxy_datasets	r
replication	r
roots	g
tests	u
.gitignore	s
LICENSE	l
README.md	u
pytest.ini	/
setup.py	bump version for new release 2 months ago

```python
from galaxy_datasets.pytorch import GZ2

gz2_dataset = GZ2(
    root='your_data_folder/gz2',
    train=True,
    download=False
)
batch = gz2_dataset[0]
image = batch['image']
label = batch['smooth-or-featured-gz2_smooth']
```

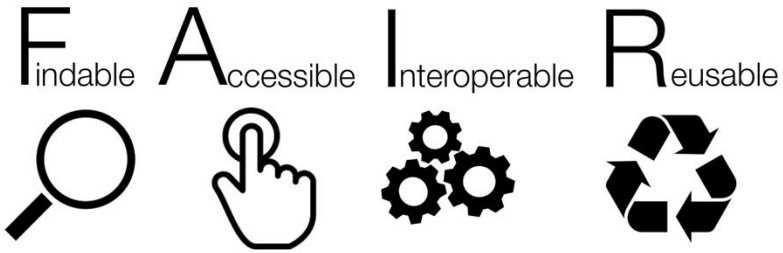
In [42]:

```python
top_5_predictions = predictions.sort_values('ring_pred', ascending=False)
show_rings(top_5_predictions)
```

https://github.com/mwalmsley/galaxy-datasets

Is your data ready for AI?

A special flavor of FAIR



Addressing sufficiently general questions?

Python access? Other data sources?

How can non-experts understand the question?

Is the data visible?

Simulation and Data Lab Applied Machine Learning

Consulting

Data	Collect, select, transform, process
Question	Formulate, refine, transform
Method	Select, develop, modify, improve
Implementation	Create, develop, optimize, parallelize
Communication	Write, speak, present, reflect

WEST AI KI-Servicezentrum

contact@westai.de

Bundesministerium für Bildung und Forschung

UNIVERSITÄT BONN

JÜLICH Forschungszentrum

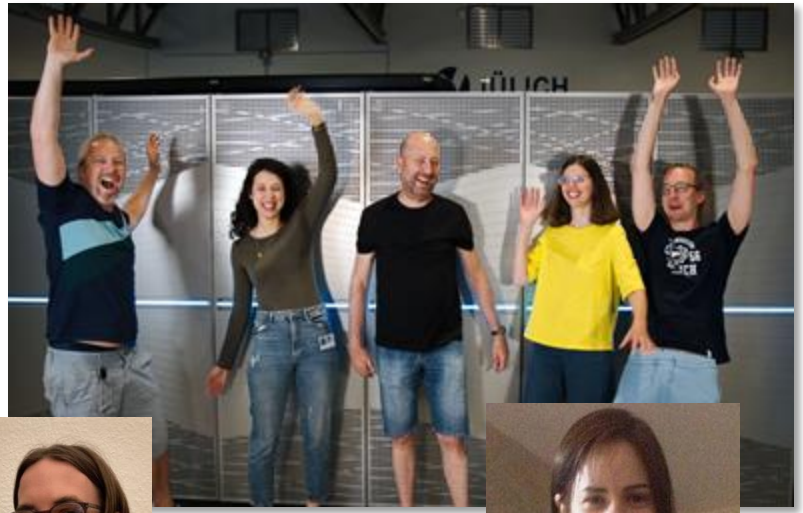
RWTH AACHEN UNIVERSITY

HELMHOLTZ AI | ARTIFICIAL INTELLIGENCE COOPERATION UNIT

https://www.helmholtz.ai

Simulation and Data Lab Applied Machine Learning

SDLAML – AI for science and more!



s.kesselheim@fz-juelich.de

LLM Applications in High Energy Astronomy

Introducing ‘CTAgent’

Elisa Jones
30.09.2025

HELMHOLTZ

DESY.

Overview

The need for correct code

Models used:

Qwen3 30B A3B

GPT-OSS

Streamline for Data Modelling

- Data modelling is skill-specific and time consuming
- Modern telescopes involve vast datasets and complex configurations
- Pydantic model generator
- Uses a feedback agent
- Data safety & Control from open source, controllable environment

Microsoft AutoGen

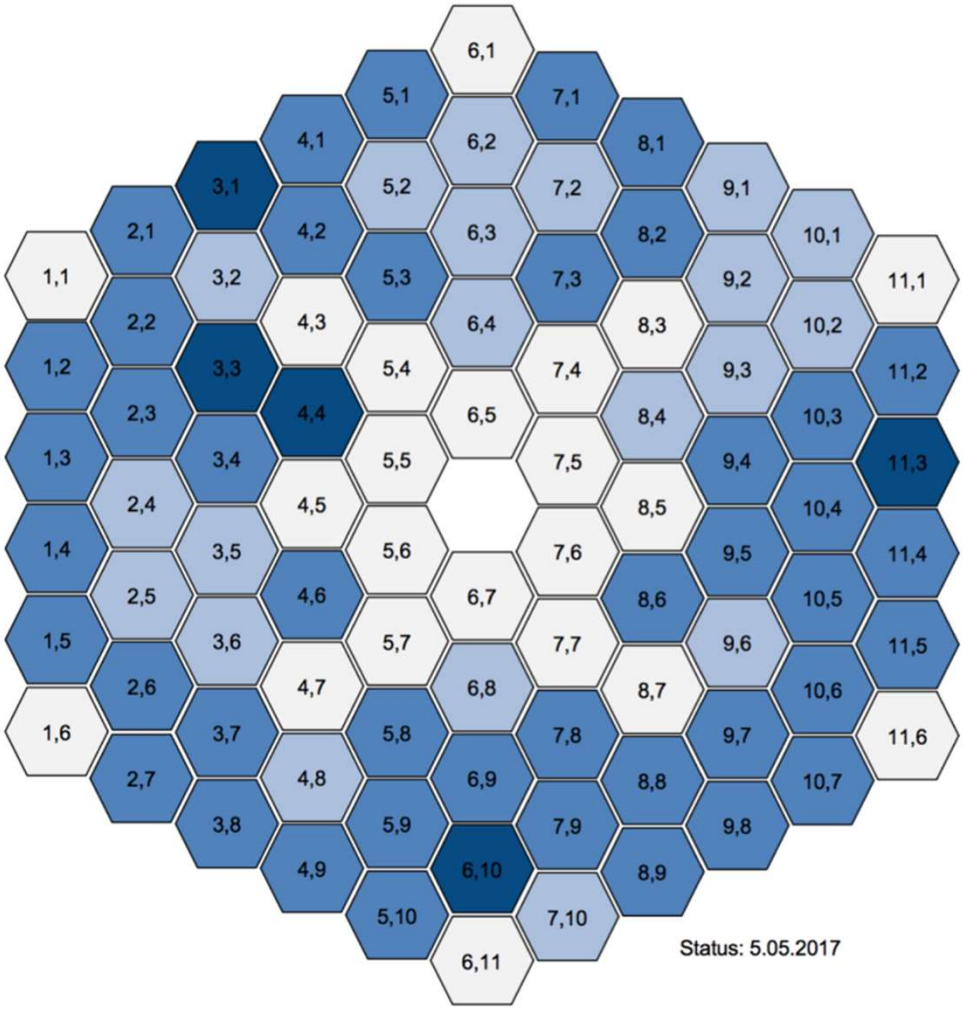
- Multiagent Framework
- AssistantAgent allows to configure prompts, API keys, and behaviour
- Tool registration
- Controls workflow

Pydantic Models

In the context of Telescope Configuration

- Ensure the many different configuration parameters are valid before use
- Hierarchy represented intuitively through nesting
- Display data types for validation
- Easy exporting back to JSON for logging

```
class DriveSystemConstants(BaseModel):
    longitude_value_degrees: float
    latitude_value_degrees: float
    altitude_value_meters: float
    altitude_zero_point_degrees: float
    altitude_maximum_degrees: float
    altitude_minimum_degrees: float
    azimuth_zero_point_degrees: float
    azimuth_maximum_degrees: float
    azimuth_minimum_degrees: float
    temperature_value_celsius: float
    pressure_value_hPa: float
    humidity_value_percent: float
    eff_wavelength_value_nm: float
```



Cherenkov Telescope Array Observatory. (2024, September 25). *CTA MST structure configuration* (Version 1.3, MST-STR-ICD-36141000-00008). [Internal document]

AutoGen Features

A framework for Multiagent systems with AgentChat

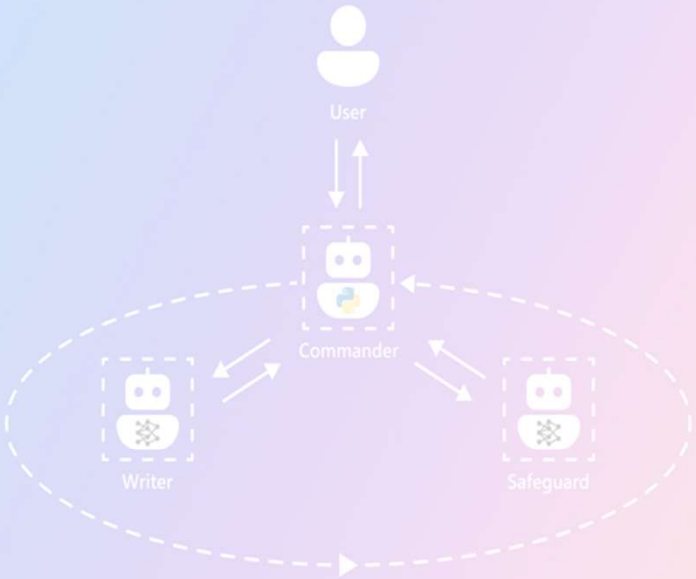
AssistantAgent

- Built-in agent with the ability to use tools
- CTAgent App uses four instances of AssistantAgent, each configurable by prompts
- Function for emitting only python code, assisted by ‘CodeImprover’ Agent

Function Tools

- ‘emit_python’ tool is registered to each code agent.
- Final validation for outputting pure code

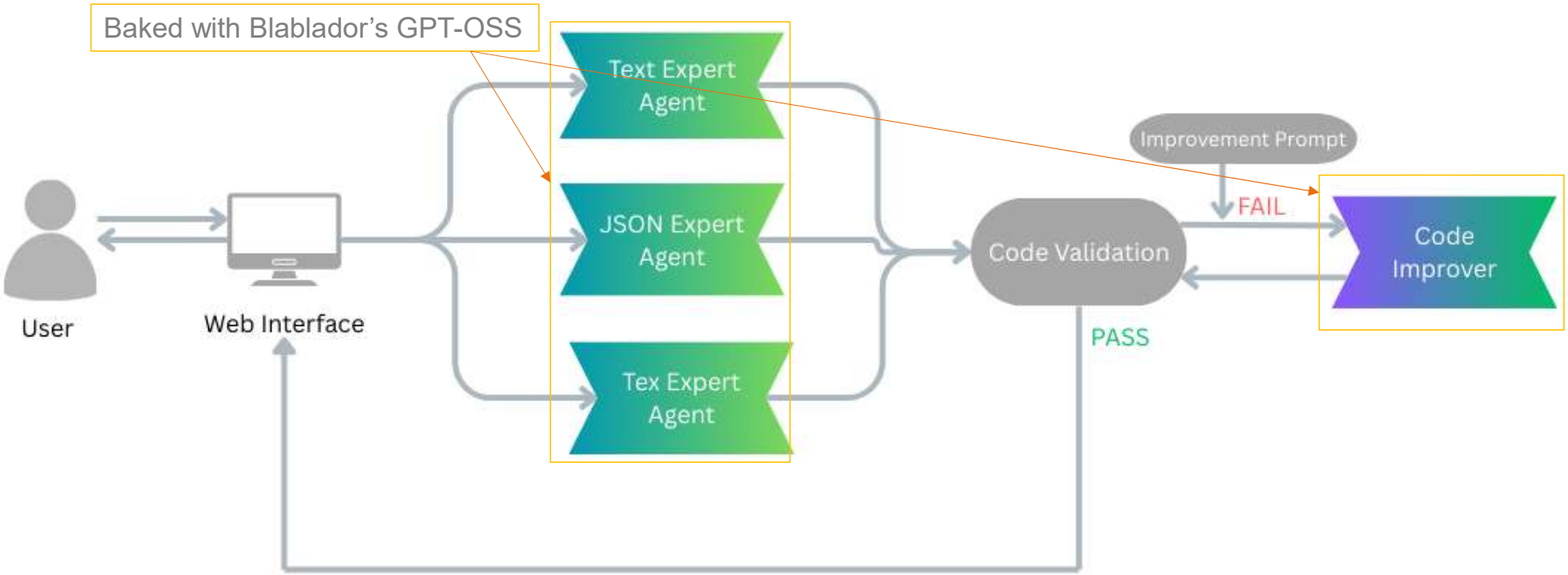
Image by Microsoft
Autogen. Microsoft Research. (2025, May 12)
https://www.microsoft.com/en-us/research/project/autogen/



Agent Workflow

Structure pipeline

JSONExpert
"…create a complete, working Pydantic BaseModel class. Preserve field names and handle nested objects, arrays, and optional fields…"



MST Configuration File

As a .tex

Ingesting file: configNote.tex | MIME: text/x-tex | Size: 46046 bytes

Detected as TEX

INFO ctagent.agents: Registered tools with all expert agents

INFO ctagent.agents: Starting pipeline with content length: 46038, type: txt

INFO ctagent.agents: Selected expert: TexExpert

INFO ctagent.agents: Attempt 1/3

WARNING ctagent.agents: No code generated on attempt 1

INFO ctagent.agents: Code generation successful on attempt 2

INFO ctagent.chat: Pipeline result: {'status': 'success', 'code':

…

Deployment

- Tested and passed on large MST pdf
- CTAgent as a web link for CTAO

```python
from typing import List, Optional
from enum import Enum

from pydantic import BaseModel, Field, IPvAnyAddress, conint, confloat, conlist


# ---------------------------------------------------------------------------- #
# Enumerations
# ---------------------------------------------------------------------------- #
class MonitoringAction(str, Enum):
    DO_NOTHING = "DO_NOTHING"
    ENABLE_NOW_DISABLE_ATEND = "ENABLE_NOW_DISABLE_ATEND"
    DISABLE_NOW_ENABLE_ATEND = "DISABLE_NOW_ENABLE_ATEND"
    ENABLE_NOW = "ENABLE_NOW"
    DISABLE_NOW = "DISABLE_NOW"


class PCAMFlag(str, Enum):
    SINGLE_SHOT = "SINGLE_SHOT"
    CONTINUOUS = "CONTINUOUS"


# ---------------------------------------------------------------------------- #
# Basic building blocks
# ---------------------------------------------------------------------------- #
class Mirror(BaseModel):
    """Global mirror list entry."""
    mirror_id: str = Field(..., alias="mirror ID")
    type: Optional[str] = None
    manufacturer: Optional[str] = None
    production_date: Optional[str] = None  # ISO-date string
    state: Optional[str] = None  # could be an enum in a real implementation


class Actuator(BaseModel):
    """Global actuator list entry."""
    actuator_id: str = Field(..., alias="actuator ID")
    address: Optional[int] = None  # hardware address, stored as integer/byte
    type: Optional[str] = None
    status: Optional[str] = None


class DeviceServer(BaseModel):
    """AMC device server (single IP address, optional ID)."""
    device_server_id: Optional[int] = None
    ip: IPvAnyAddress = Field(..., alias="IP")
```

etc…

CTAgent

Upload a text file. A Pydantic model will be generated from it.
You can also describe your model or add further instructions in the text box below.

Additional instructions or description:

```python
from __future__ import annotations

from typing import List, Literal

from pydantic import BaseModel, Field


class ActuatorConfig(BaseModel):
    actuator_id: str = Field(..., description="Unique identifier of the actuator")
    type: str = Field(..., description="Type of the actuator")
    address: str = Field(..., description="Network or physical address of the actuator")
    status: Literal["functional", "broken"] = Field(..., description="Operational status of the actuator")


class ActuatorConfigList(BaseModel):
    """
    Represents the top-level JSON array defined in the schema.
    """
    __root__: List[ActuatorConfig]
```

Drop a file or click to upload
451.0B / 100.00%

amc_cfg.json
451.0B / 100.00%

GENERATE MODEL

Conclusion

Scalable

- Agents can be added in parallel
- Function tools can be added and removed
- AutoGen framework allows for a modular pipeline

Specialized

- Agents are single-task oriented, and validated as such
- Python parsing built in
- Open-source (gpt-oss) allows for cost and energy efficient deployment

Integration

- Code-first
- Models called through external APIs
- BLABLADOR was easily accessible and it's gpt-oss was the best performing model

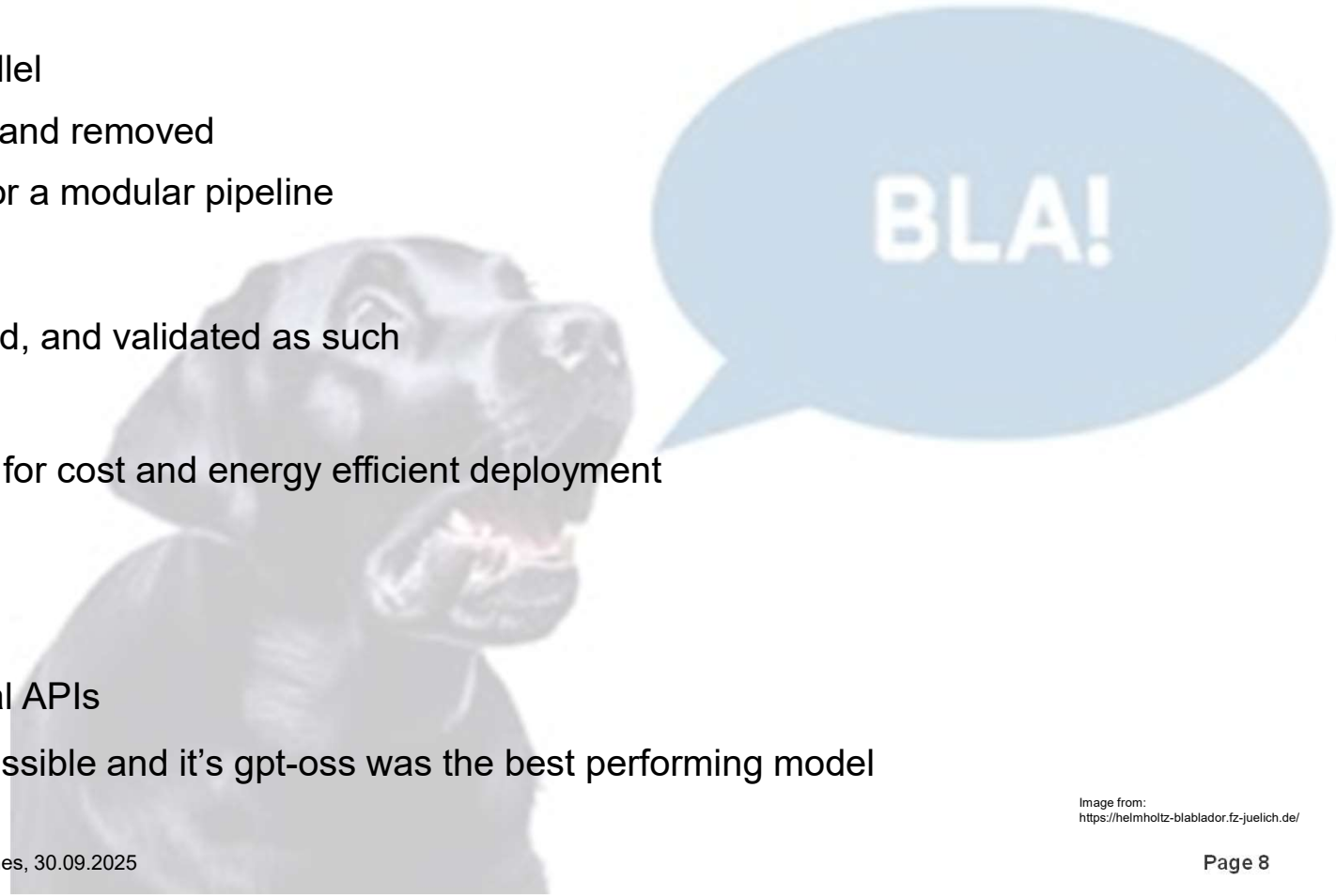


Image from:
https://helmholtz-blablador.fz-juelich.de/