EVE:

A Suite of LLM and Data for EO and Earth Sciences

Àlex R. Atrio
Antonio Lopez
Marcello Politi
Vijayasri Iyer
Christopher Phillips

Elena Christodoulou
Cristiano De Nobili
Sébastien Bratières
Nicolas Longépé

International workshop on AI Foundation Model for EO

5-7 May 2025 | ESA ESRIN - Frascati, Italy

Contents

The EVE Project

EVE: A Domain-Specific LLM and Datasets for Earth Observation

Base and instructed LLMs for Earth Observation (EO) and Earth Sciences (ES)

- Trained on billions of high-quality EO/ES data
- Natural Language Question-Answering (QA)
- Scientific summarization
- Structured data extraction from raw text
- RAG with scientific citations
- Designed for both scientific experts and non-specialist users

Open Science Contributions from EVE

- Domain-adapted and instruction-tuned LLMs
- Curated EO/ES continued pre-training corpus
- Instruction and evaluation datasets in the EO domain
- Publicly accessible chat interface hosted by ESA Φ-lab
- Legal and compliance documentation for open use

Consortium and Timeline

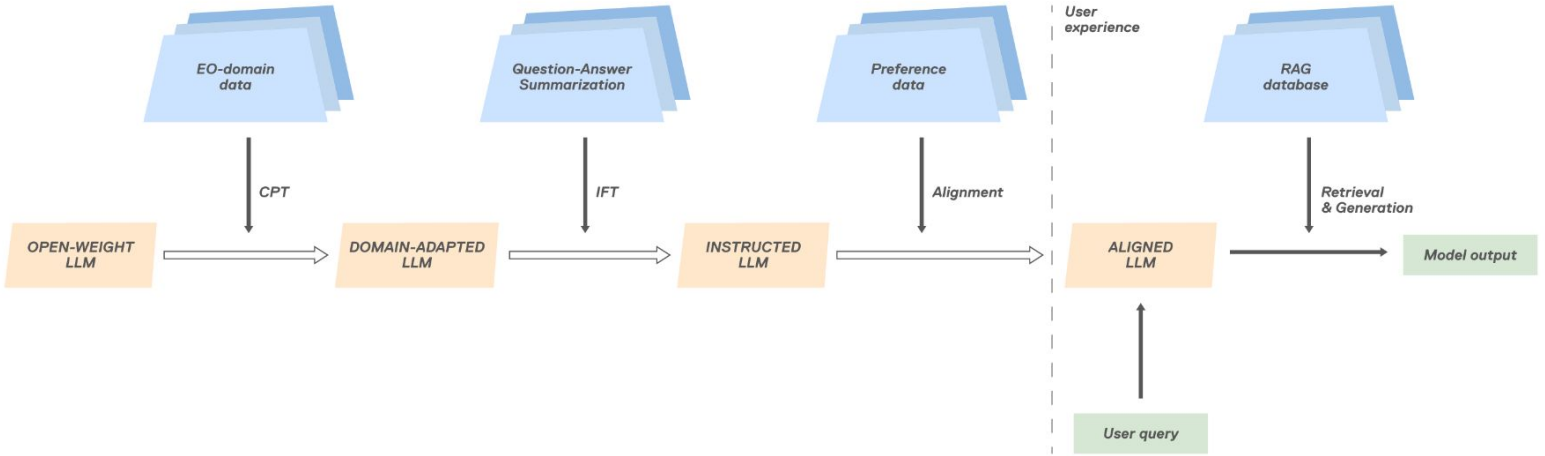
Consortium:

- Pi School (Italy)
- Imperative Space (United Kingdom / Italy)
- ESA Φ-lab – Scientific Lead: Nicolas Longépé

Timeline:

- Project Start: May 2024
- Project End: November 2025
- Pilot Phase: September – October 2025

EVE Training and Deployment Pipeline



Project Status and Next Steps

Current status

- Instructed Llama 3.1 8B on synthetic EO data (QA and summarization)
- Full ~5B token clean EO/ES corpus to be released as open-source
- RAG prototype with small subset of EO/ES data

Next Steps:

- Manual annotation/creation of EO IFT & evaluation datasets
- Perform continued pre-training on raw EO/ES text
- Expand IFT with broader, higher-quality data
- Scale up to larger architecture (20B-70B range)

Motivation

End-to-End Project

EVE is not just a model — it's a full suite including:

- Curated EO/ES corpus (raw)
- Curated EO/ES instruction datasets
- Curated EO/ES benchmarking datasets
- Training and RAG-creation workflows
- RAG built on high-quality scientific databases
- Dedicated hallucination detection pipeline

Model Motivation

- Domain-specific LLMs often match generalist models on technical domains
  - (e.g., K2, CosmoSage, PMC-LLaMA, AstroMLab)
- Small model (<70B) means more affordable deployment and use
- Open weights (quantizable, fine-tunable, etc.)
- Manual evaluation on real EO scenarios to demonstrate utility

Data

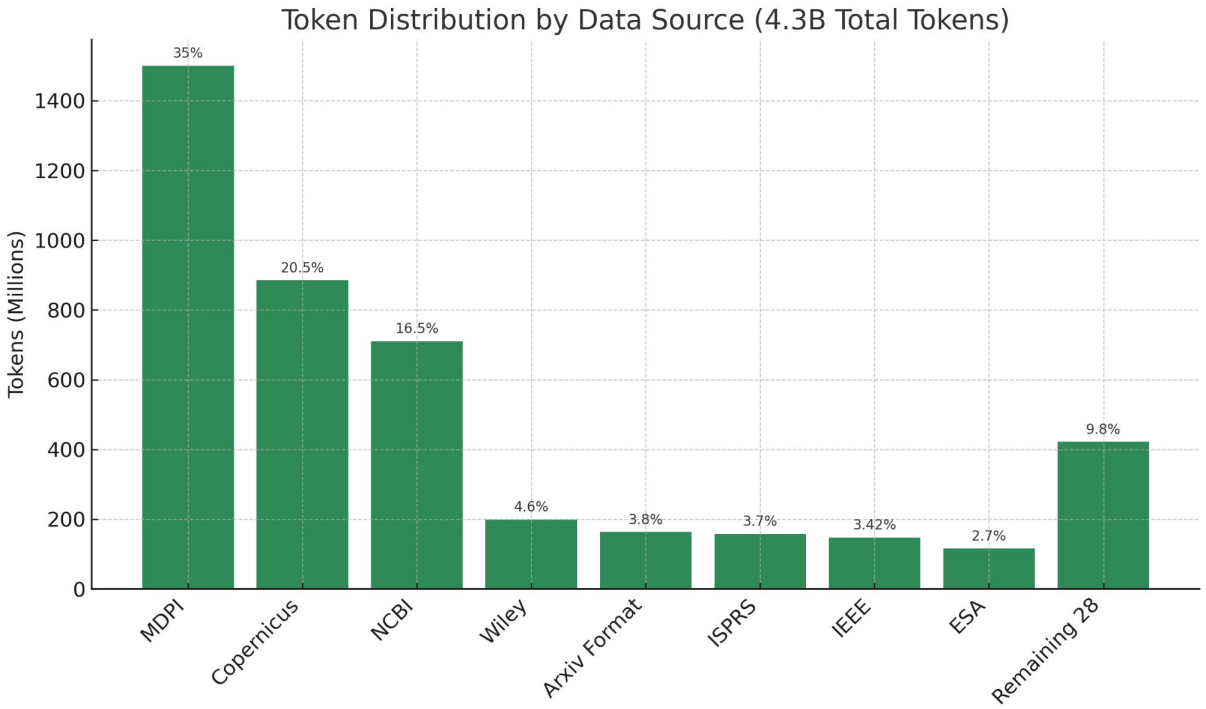
Overview of EO data

Data	Quantity of data	Type of data
Scientific EO data	**5.7B** tokens Open-Access **1.1B** tokens Closed-Access (Wiley)	Raw data
Authentic SFT data	**5k** QA samples **>100k** Summarization samples	Instruction, input, output
Synthetic SFT data	**>1M** samples	Instruction, input, output
Preference	**5K** samples	Prompt, winning_response, losing_response
RAG database	**300k** articles/documents	Indexed chunks of text

Corpus Creation Workflow

- Manual identification and scraping of high-quality, trusted sources
- Custom extraction and processing pipeline
- Metadata enrichment (journal, year, etc.)

CPT Corpus Data Breakdown



Instruction Fine Tuning Data

- Authentic generation:
  - 5k samples for evaluation and synthetic generation
- Synthetic generation:
  - Grounded LLM-based generation on multi-chunk document context

Model

CPT

- Full precision Causal Language Modeling on corpus described above
- In addition to EO: math & code to improve reasoning abilities
- Annealing
  - Final phase cycling learning rate while upsampling high-quality data

IFT, Alignment, Model Merging

- IFT
  - QA
    - Open-ended
    - MCQA
    - With/without context
  - Summarization
  - Structured data extraction from raw text
- Alignment
  - Authentic preference EO dataset
- Model Merging
  - With generalist instructed LLM allows to mitigate catastrophic forgetting

RAG and factuality

Hallucination/factuality

- Objective: factuality-based hallucinations
  - Science data CPT decreases hallucinations in science and general domain (Li et al., 2024)
- Workflow:
  - Flagging: Uncertainty (internal LLM probability) / LLM-as-a-judge
  - Re-prompting (cf. Chain of Verification) a fact-checker LLM with RAG and trusted online sources connection
  - Newly found conflicting data added to RAG db
  - EVE re-prompted with new data and to focus on problematic spans of text
- Evaluate workflow on our manually created EO hallucinations dataset

Evaluation

Evaluation Datasets and Benchmarks

General-Domain Evaluation

- MMLU, GPQA, MATH, HaluEval (zero-/few-shot), etc.
- Selected texts: perplexity before and after fine-tuning (vs. baseline models)

EO-Specific Evaluation

- Open-ended QA and MCQA
- Hallucination/factuality dataset
- Red teaming sets (adversarial prompts for robustness testing)

All EO dataset manually created, to be open-sourced

Conclusions

Thanks!

What's Next for EVE:

- Pilot phase (Sep – Oct 2025)
- Open-sourcing (Nov 2025)
- Project end (Nov 2025)

Tomorrow, with INDUS (NASA):

- 10h15 deep dive
- 15h00 hands-on tutorial

Scan and register for the pilot