Local AI Infrastructure for Scientific Workflows

Domain-Specific LLMs on NVIDIA Grace-Blackwell

Giovanni Lacopo
Tecnologo a tempo determinato — INAF OATS
giovanni.lacopo@inaf.it
Collaborators: H. Heinl, G. Taffoni, A. Ragagnin, D. Goz

SAIt 2026 • L'Aquila • May 8, 2026

The Challenge

Next-generation surveys (e.g. SKA, Euclid) generate exponentially growing datasets that exceed traditional analysis paradigms.

Data Sovereignty

Keep sensitive research data on-premises, fully controlled

Domain Expertise

Science-specific models outperform general LLMs

Cost Efficiency

Local inference at a fraction of commercial API costs

GB10 Grace-Blackwell Infrastructure

Hardware Specifications

- 5 × Lenovo PGX Spark GB10 nodes
- 128 GB unified memory per node (~111 GB usable)
- ARM64 Grace CPU + Blackwell GPU (sm_121)
- InfiniBand 100 Gbit/s (2 ports available, 1 active)
- vLLM inference via NVIDIA NGC (26.03)
- Total: ~555 GB inference capacity

111 GB

GPU memory per node

~273 GB/s

memory bandwidth

FP8

quantization (all models)

100 Gbit/s

InfiniBand (1 active port)

Deployed Models

Qwen3-Coder-Next

Agentic Coding

- MoE: 80B / 3B active
- FP8 quantized
- 45K context
- Tool calling enabled
- ~90 tok/s

Qwen3.5-35B-A3B

General Assistant

- MoE: 35B / 3B active
- Reasoning (thinking)
- 131K context
- Tool calling enabled
- ~90 tok/s

Qwen3.5-122B-A10B

Orchestrator

- MoE: 122B / 10B active
- Reasoning (thinking)
- 131K context
- 2-node PP over IB
- ~25 tok/s

Qwen3-Next-80B

Documentation

- MoE: 80B / 3B active
- Instruct variant
- 131K context
- FP8 quantized
- ~90 tok/s

Multi-Node Deployment: Qwen3.5-122B

Pipeline Parallelism over InfiniBand — distributing 122B parameters across 2 nodes

Node A — Layers 1–30

First half of model layers
~61 GB weights (FP8)
Receives user input
Passes activations to Node B

InfiniBand
100 Gbit/s

Node B — Layers 31–60

Second half of model layers
~61 GB weights (FP8)
Receives activations from A
Produces final output

Why PP instead of TP? Tensor Parallelism (TP) splits each layer across GPUs, requiring an all-reduce synchronization at every layer (~40×/token). Pipeline Parallelism (PP) splits the model in two halves: layers 1–30 on Node A, layers 31–60 on Node B. Activations are passed only once between stages. With InfiniBand at ~12 GB/s (vs NVLink at 600+ GB/s), PP minimizes communication overhead.

Multi-Agent Coding Architecture

Seven specialized agents powered by different LLMs, coordinated via OpenCode

Orchestrator — Qwen3.5-122B (10B active, reasoning)

Researcher
Qwen3.5-35B

Senior Coder
Qwen3.5-122B

Coder
Qwen3-Coder

Writer
Qwen3.5-35B

Executor
Qwen3.5-35B

Scrum Master
Qwen3.5-35B

Workflow: User → Orchestrator → delegates to Researcher (web docs) → Senior Coder / Coder (implementation) → Executor (compile & test) → Writer (documentation). Scrum Master monitors and escalates.

Performance Highlights

Model	Total Params	Active	Throughput	Context	Nodes
Qwen3-Coder-Next	80B (MoE)	3B	~90 tok/s	45K	1
Qwen3.5-35B-A3B	35B (MoE)	3B	~90 tok/s	131K	1
Qwen3-Next-80B	80B (MoE)	3B	~90 tok/s	131K	1
Qwen3.5-122B-A10B	122B (MoE)	10B	~25 tok/s	131K	2 (PP)

Cost: Local vs Cloud API

1M tokens local: ~$0
(electricity only)
1M tokens GPT-4o API: ~$15

24/7 availability • No API rate limits • Full data privacy • Reproducible

Next Step: ScienceQwen Fine-Tuning

Combining MoE efficiency with multi-disciplinary scientific expertise

Astrophysics
~280K papers

Physics
~250K papers

Mathematics
~200K papers

Chemistry &
Biology ~270K

~1M

arXiv papers total

122B / 10B

total / active params

LoRA

parameter-efficient FT

Goal: a model that reasons across physics, mathematics, chemistry, and biology — not just astrophysics

Training on Leonardo

Compute Resources

- Leonardo supercomputer (CINECA)
- 256 nodes × 4 A100 64GB = 1,024 GPUs
- 256 GB HBM2e per node
- InfiniBand HDR fabric
- ~60,000 A100-hours estimated
- 7-10 days wall-clock time

Training Pipeline

1. Continued Pre-Training

2–3 epochs on ~1M arXiv papers
(astro-ph + physics + math + chem + bio)

2. Supervised Fine-Tuning

Synthetic Q&A from domain experts
LoRA adapters (rank 64) on attention layers

Why LoRA instead of full fine-tuning?

LoRA trains only small adapter matrices (~1–2% of params), reducing GPU memory by ~10× and preserving the base model's general knowledge.

Expected Benefits

~7× Faster Inference

MoE activation (10B) vs dense 70B models

Multi-Discipline Expertise

Physics, math, chemistry, biology — not just astrophysics

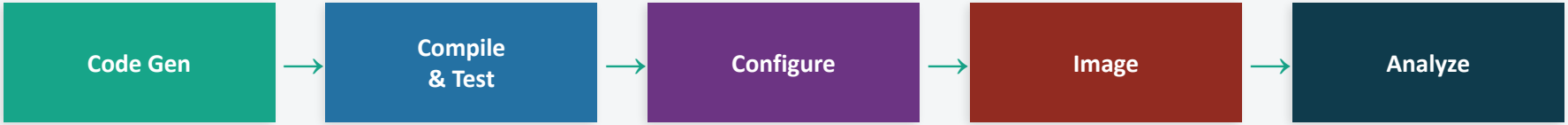
Reasoning + Knowledge

Thinking mode solves complex scientific problems step by step

Long-Context Mastery: up to 131K tokens = ~100 pages of documentation per session

Vision: Agentic AI for SKA Pipelines

Automated data reduction pipelines: imaging → source extraction with RICK



LLM Agent Capabilities

- Automated RICK code compilation and HPC optimization (OpenMP, MPI, CUDA)
- Parameter tuning and execution monitoring on GB10 cluster
- Results analysis and quality validation with reasoning models
- Integration with RICK imaging kernels (Lacopo+ 2026, doi:10.1016/j.ascom.2026.101074)
- Multi-agent workflow: researcher → coder → executor → writer (full pipeline)

Key Takeaways

- Local AI infrastructure enables data sovereignty + cost efficiency for scientific research
- MoE architectures (3–10B active) deliver high throughput on compact hardware (GB10)
- Multi-node pipeline parallelism scales 122B-parameter models over InfiniBand
- Multi-agent workflows automate the full scientific coding pipeline
- ScienceQwen will bring multi-disciplinary domain expertise (physics, math, chemistry, biology)
- Agentic AI will transform astronomical data analysis for SKA and beyond

Thank You

Giovanni Lacopo
giovanni.lacopo@inaf.it
INAF — Osservatorio Astronomico di Trieste

SAIt 2026 • L'Aquila • May 8, 2026