Local AI Infrastructure for Astronomical Workflows

Domain-Specific LLMs on NVIDIA Grace-Blackwell

Giovanni Lacopo
Tecnologo a tempo determinato presso INAF OATS
giovanni.lacopo@inaf.it

Collaborators: Hendrik Heinl, Giuliano Taffoni, Antonio Ragagnin, David Goz

USC-C General Assembly • Trieste • March 10, 2026

The Challenge

Next-generation surveys **(i.e. SKA)** are generating **exponentially growing datasets** that exceed traditional analysis paradigms.

Data Sovereignty

Keep sensitive research data on-premises

Domain Expertise

Astronomy-specific models outperform general LLMs

Cost Efficiency

Local inference at fraction of API costs

GB10 Grace-Blackwell Infrastructure

Hardware Specifications

- 5 NVIDIA GB10 Grace-Blackwell nodes
- 128 GB unified memory per node
- ARM64 Grace CPU + Blackwell GPU
- Ethernet 10 Gigabit connection
- Total: ~550 GB inference capacity

Key Features

110 GB
GPU memory per node

80
tokens/second

FP8 (~2PFlops)
quantization support

Deployed Models

AstroSage-8B

Domain Reasoning

- Llama 3.1-8B base
- de Haan+ 2025 fine-tuning
- ~80% AstroMLab (8B)
- 70B: 86.2% (beats GPT-4)
- 10-12 tok/s
- 65k context

Qwen3-Coder-80B

Agentic Coding

- MoE architecture
- 80B / 3.9B active
- FP8 quantized
- Multi-language
- ~80 tok/s
- Ideal for agentic workflows

Qwen3-Next-80B

Documentation

- MoE architecture
- 80B / 3.9B active
- FP8 quantized
- 262k context
- ~80 tok/s
- Tech manuals

Performance Highlights

Model	Parameters	Active	Throughput	Memory
AstroSage-8B	8B	8B	10-12 tok/s	16 GB
Qwen3-Coder-80B	80B	3.9B (MoE)	~80 tok/s	80 GB
Qwen3-Next-80B	80B	3.9B (MoE)	~80 tok/s	80 GB
GPT-4o (API)	~1.8T	Unknown	~50 tok/s	Cloud

Cost: Local vs API

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

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

Next Step:

AstroQwen Fine-Tuning Project

Combining the best of both worlds: MoE efficiency + astronomy expertise

288k+	80B / 3.9B	~80 tok/s
arXiv papers	Total / Active params	Target throughput

Training on Marco Polo

Compute Resources

- Marco Polo supercomputer
- 64 nodes × 4 H100 = 256 GPUs
- 320 GB HBM3 per node
- InfiniBand HDR/NDR fabric
- ~53,000 H100-hours total
- 8-10 days wall-clock time

Training Pipeline

1. Continued Pre-Training

2.5 epochs on astro-ph + textbooks

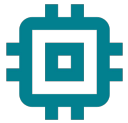
2. Supervised Fine-Tuning

Synthetic Q&A

Expected Benefits

20× Faster Inference

MoE activation (3.9B) vs dense 70B models



Superior Performance

Match AstroSage-70B quality with speed

Long-Context Mastery

262k tokens = ~200 pages (full manuals)

Vision: Agentic AI for SKA Pipelines

Automated data reduction pipelines: **imaging → source extraction with RICK (Radio Imaging Code Kernels)**

LLM Agent Features:

- Automated RICK code compilation and optimization
- Parameter tuning and execution monitoring
- Results analysis and quality validation
- Integration with RICK imaging kernels (Lacopo+ 2026, https://doi.org/10.1016/j.ascom.2026.101074)

Vision: Agentic AI for SKA Pipelines

Image produced by NanoBanana Pro

Key Takeaways

- Local AI infrastructure enables data sovereignty + cost efficiency
- Domain-specific models outperform general LLMs on astronomy tasks
- MoE architecture provides 20× speed improvement over dense models
- Agentic workflows will transform astronomical data analysis
- AstroQwen combines efficiency with state-of-the-art performance

Thank You

Giovanni Lacopo
giovanni.lacopo@inaf.it
INAF OATS • Trieste