ORION: A Supervisory Architecture For AI-Assisted Autonomous Beamline Operation

SERGIO LOPEZ-CACERES
Postdoctoral Appointee
Physics Division
Argonne National Laboratory

DEDICATED AI/ML ACCELERATOR TEST FACILITY WORKSHOP
July 15-17, 2026
SLAC National Accelerator Laboratory
Menlo Park, CA, USA

Argonne NATIONAL LABORATORY

OUTLINE

Part 1

Application of ML to radioactive ion beam tunning

Part 2

Development of a custom multi-section optimization tool

PART 1 Application of ML to radioactive ion beam tunning

Argonne NATIONAL LABORATORY

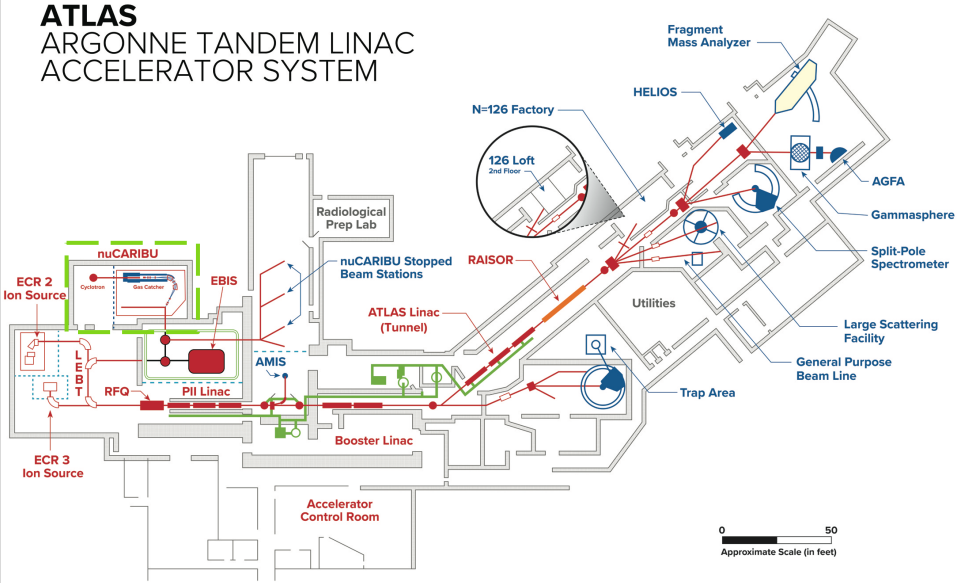
WHAT IS nuCARIBU? CALIFORNIUM RARE ISOTOPE BREEDER UPGRADE

A radioactive ion source part of the ATLAS national user facility

Characteristics

- Research in **nuclear physics, nuclear astrophysics, and national security**
- **nuCARIBU** provides beams of heavy ions made from neutron-induced fission of actinide targets
- Radioactive heavy ions beams
- Manual tunning until recently

nuCARIBU beamline

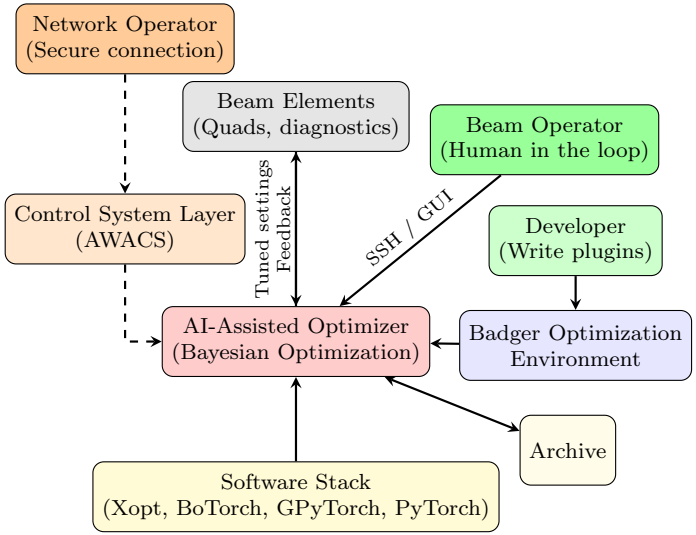


nuCARIBU + BO FIRST STEPS

Software stack

- Implemented inside **Badger*** optimization environment
- Backend: **Xopt** → **BoTorch** → **GPyTorch** → **PyTorch**
- **Human-in-the-loop** via Badger GUI
- Per-run results archived for reproducibility

Optimization framework

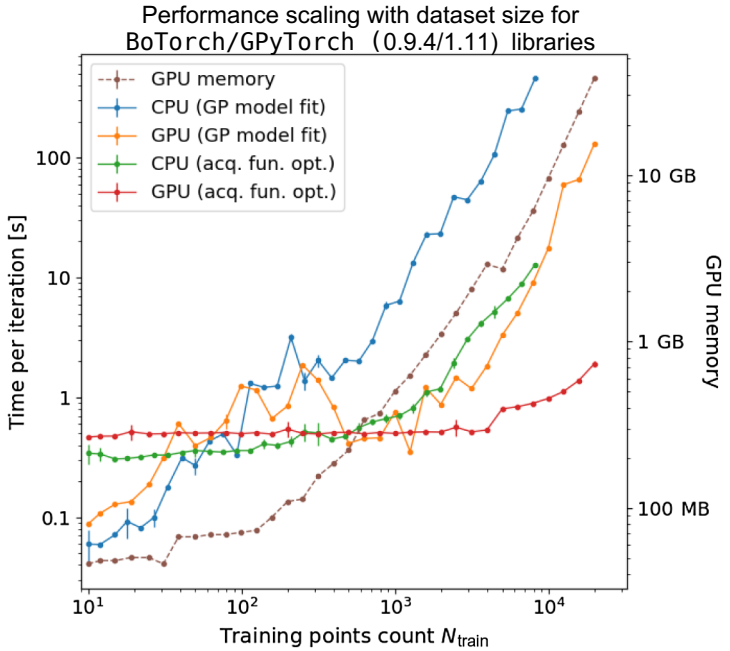


* https://github.com/xopt-org/Badger
Zhang, Z., et al. "Badger: The missing optimizer in ACR", in Proc. IPAC'22, Bangkok.

OPERATIONAL DEPLOYMENT

Computational scaling

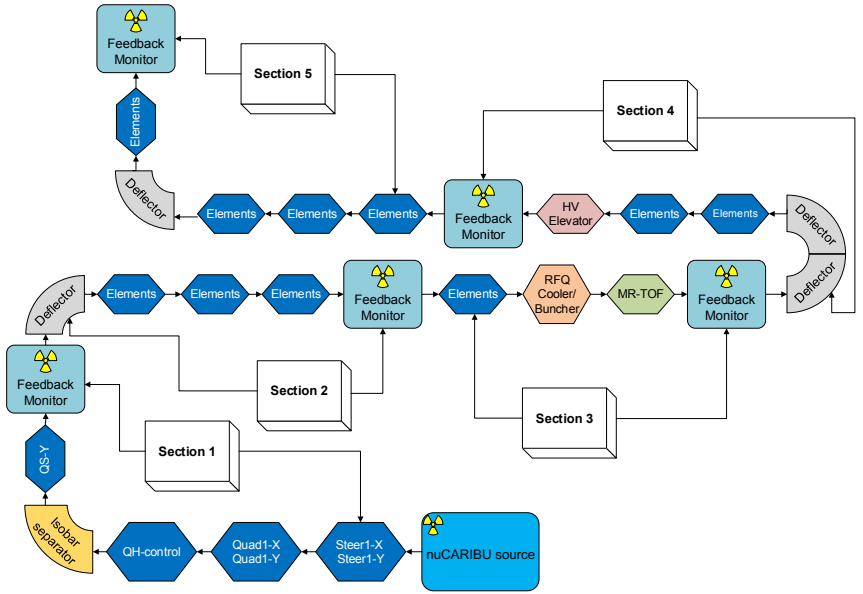
BO scaling $\mathcal{O}(n^3)$ with dimension can become a bottleneck



Roussel et al. PRAB 27, 084801 (2024)

Sectional tuning strategy

Section-by-section optimization (~2-9 controls each)

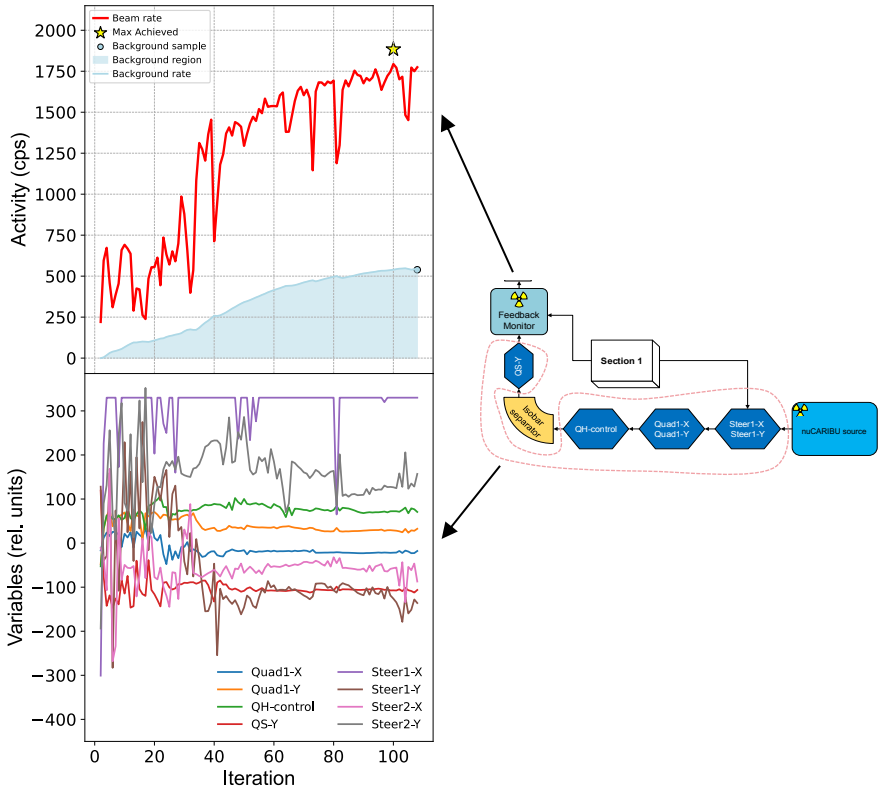


RESULTS

Transmission

- BO iteratively **improves rate**, converges in ~ 100 iterations
- Automatically **recovers** from disturbances around iteration 70 and 80
- Each iteration took about 10 s
- Control variables sample large space up to ~ iteration 40, then focuses on promising regions, while complies with safety **constraints**
- Typical convergence: **≈15 min**, **50% total transmission**, comparable to expert tuning

Example section optimization



Lopez-Caceres et al. PRAB 28, 072802 (2025)

LESSONS LEARNED

BO works without prior data in RIB	Sectional tuning	Benefits	Gaps
BO performed robustly without historical or simulated pretraining, confirming its suitability for the first ever implementation of automated tuning at the CARIBU/nuCARIBU facility	Focusing on local observables (specific feedback monitors) and beamline elements (2-9 parameters) improved convergence by reducing workload on BO models, which are known to struggle in high-parameter spaces	• Repeatability • Automation • Reduced operator load	• Automatic beam section switching missing • No self-improvement mechanisms • No warm starts • Non-intuitive UI

PART 2

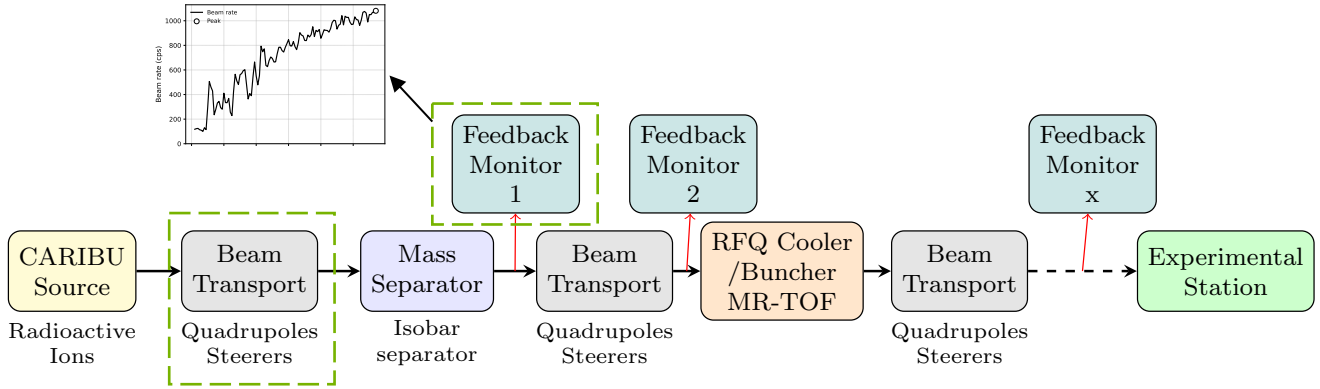
Development of a custom multi-section optimization tool

Argonne NATIONAL LABORATORY

Full automation

AI orchestration for multi-section tuning:

Delegating decision-making will be essential to achieve end-to-end automated tuning

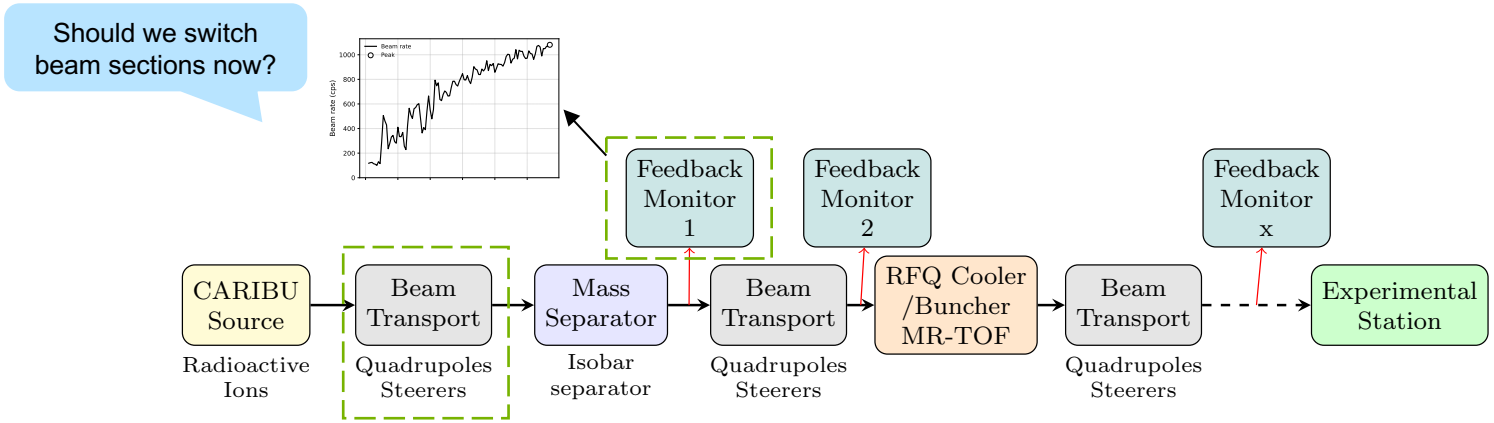


Argonne NATIONAL LABORATORY

Full automation

AI orchestration for multi-section tuning:

Delegating decision-making will be essential to achieve end-to-end automated tuning

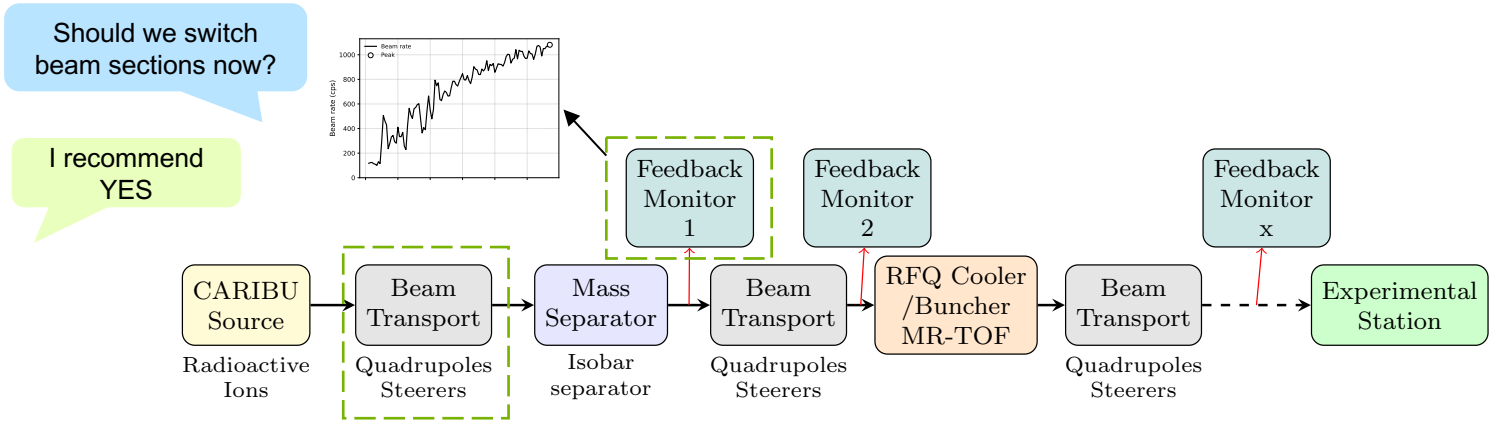


Argonne NATIONAL LABORATORY

Full automation

AI orchestration for multi-section tuning:

Delegating decision-making will be essential to achieve end-to-end automated tuning

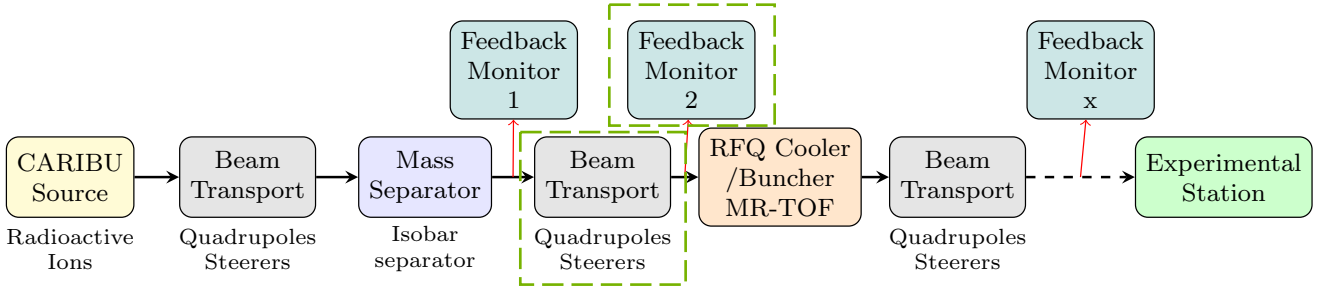


Argonne NATIONAL LABORATORY

Full automation

AI orchestration for multi-section tuning:

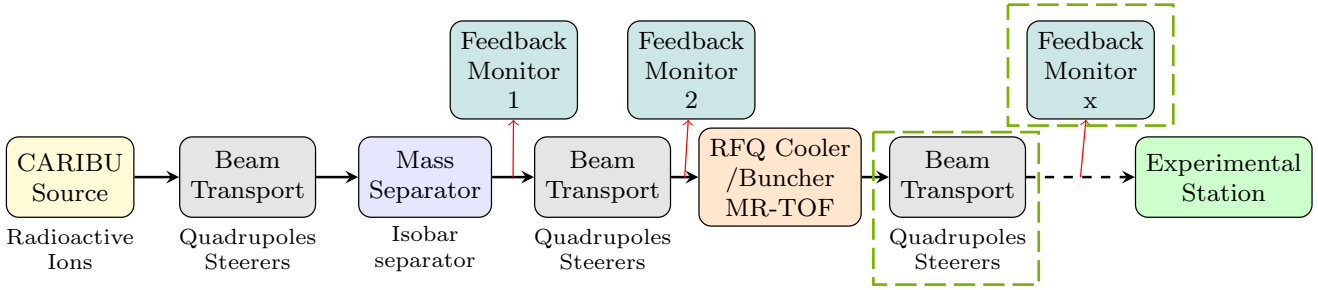
Delegating decision-making will be essential to achieve end-to-end automated tuning



Full automation

AI orchestration for multi-section tuning:

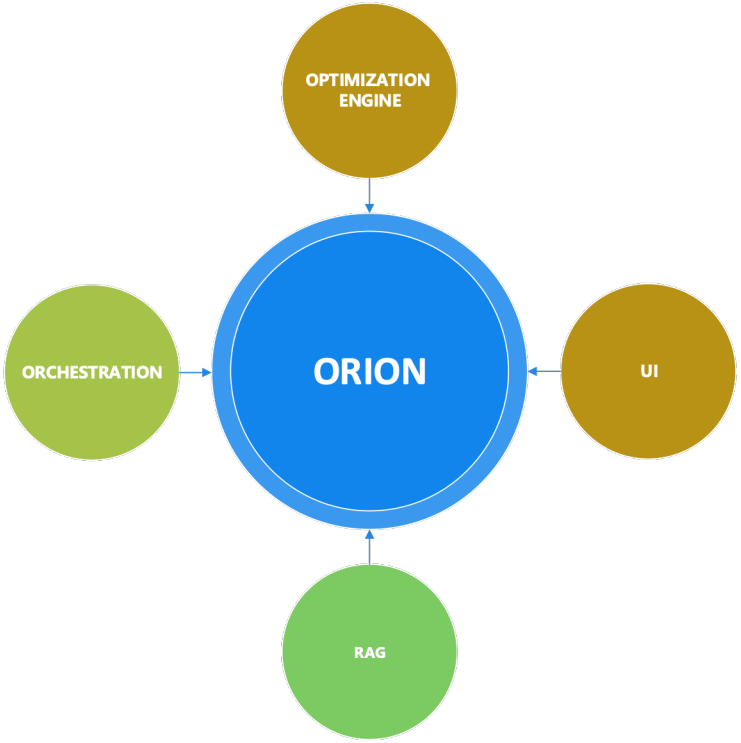
Delegating decision-making will be essential to achieve end-to-end automated tuning



THE SOLUTION

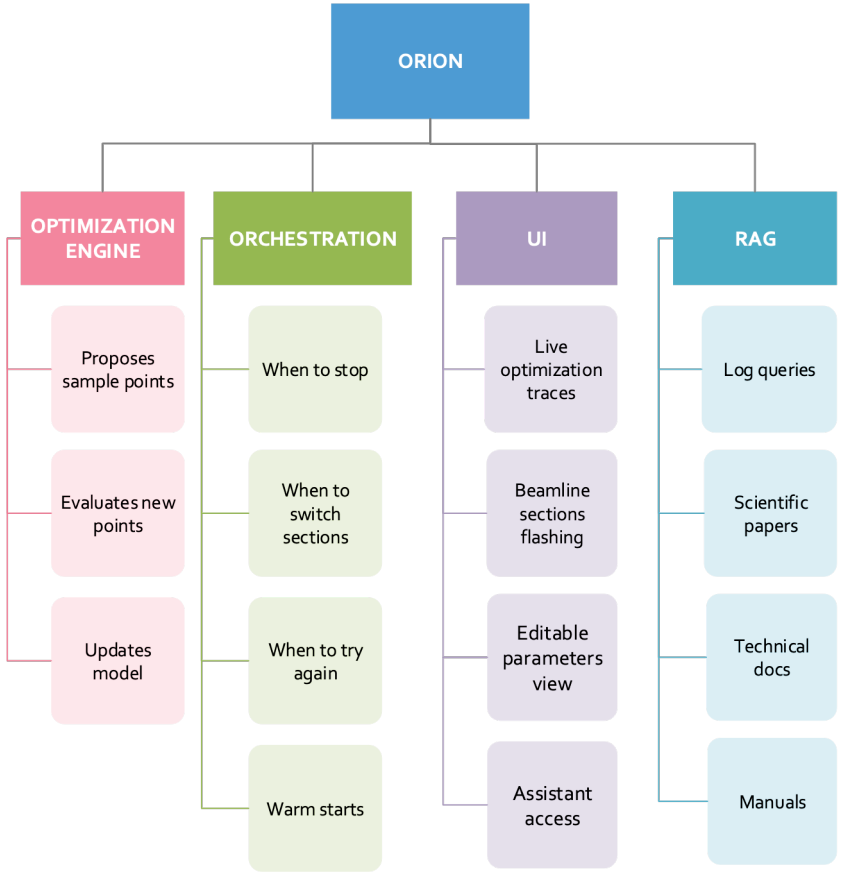
ORION: Orchestrated Real-Time Intelligent OptimizatioN

“ORION is a system that automates and orchestrates optimization across sections using Bayesian optimization + LLM-enhanced decision logic”



ORION: Orchestrated Real-Time Intelligent Optimization

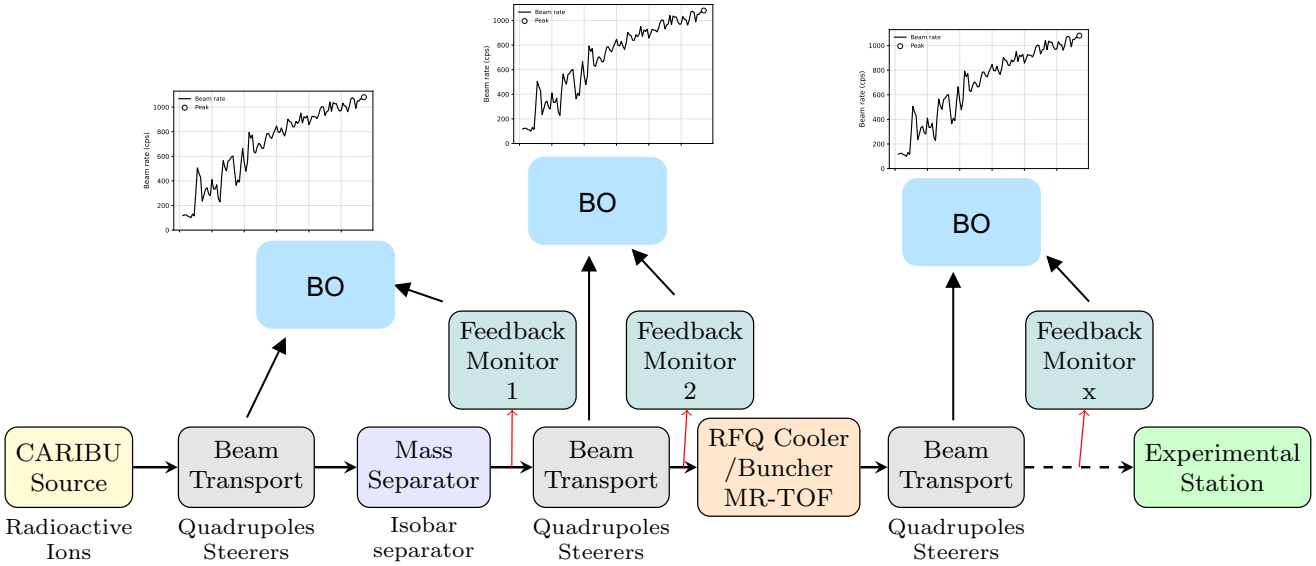
System architecture: *"Separation of concerns"*



ORION: Orchestrated Real-Time Intelligent Optimization

Bayesian Optimization Layer:

"Retain deterministic Bayesian Optimization for numerical optimization"



Argonne NATIONAL LABORATORY

ORION: Orchestrated Real-Time Intelligent Optimization

LLM-enhanced Orchestration Layer:

Current climate on frontier models usage: A shift from tokenmaxxing.

CNBC MARKETS BUSINESS INVESTING TECH POLITICS VIDEO INVESTING CLUB PRO LIVESTREAM

TECH

OpenAI and Anthropic face new AI reality as users shift from 'tokenmaxxing' to efficiency

PUBLISHED FRI, JUN 26 2026·8:00 AM EDT | UPDATED FRI, JUN 26 2026·11:33 AM EDT

electrek

TESLA

Tesla caps employee AI spending at $200/week except for Grok

Fred Lambert | Jul 2 2026 - 5:34 pm PT 48 Comments

BUSINESS INSIDER

Subscribe | Newsletters

THE GREAT CODING RESET + How AI is redefining the software engineering industry More stories

The next office power struggle: AI tokens

Workers are competing for compute as companies rethink budgets

Search FORTUNE

AI • TECH

Microsoft reports are exposing AI's real cost problem: Using the tech is more expensive than paying human employees

By Jake Angelo
Former News Fellow
May 22, 2026, 12:56 PM ET

Gizmodo LATEST NEWS REVIEWS IO9 SCIENCE DEALS DOWNLOADS NEWSLETTERS

ARTIFICIAL INTELLIGENCE

Companies Are Getting Burned by Burning Tons of Tokens

Wait, this costs money?

BY AJ DELLINGER PUBLISHED MAY 30, 2026, 7:00 AM ET

READING TIME 2 MINUTES

ORION: Orchestrated Real-Time Intelligent Optimization

LLM-enhanced Orchestration Layer:

Given the current shift from tokenmaxxing, it is worth considering local models

Advantages of Local LLM deployment

- **Offline operation.** ORION operates without internet connectivity or external API access, making it suitable for accelerator facilities with restricted or isolated networks. Lower costs.
- **Low and predictable latency.** Eliminates network delays and cloud-service variability, enabling consistent supervisory response times during optimization.
- **Data privacy and security.** Machine settings, operational logs, and facility information remain entirely within the laboratory infrastructure.
- **Independence from external services.** Operation is unaffected by internet outages, API changes, pricing models, or service availability.
- **Reproducibility.** The exact model weights, prompts, inference parameters, and software stack can be preserved, allowing experiments to be repeated under identical conditions.
- **Full control over the inference pipeline.** Prompt templates, safety checks, structured output validation, and logging can be customized for facility-specific workflows.
- **Facility-specific customization.** Local models can be paired with custom prompts, engineering heuristics, or future fine-tuning using accelerator-specific datasets.

ORION: Orchestrated Real-Time Intelligent Optimization

LLM-enhanced Orchestration Layer:

Given the current shift from tokenmaxxing, it is worth considering local models

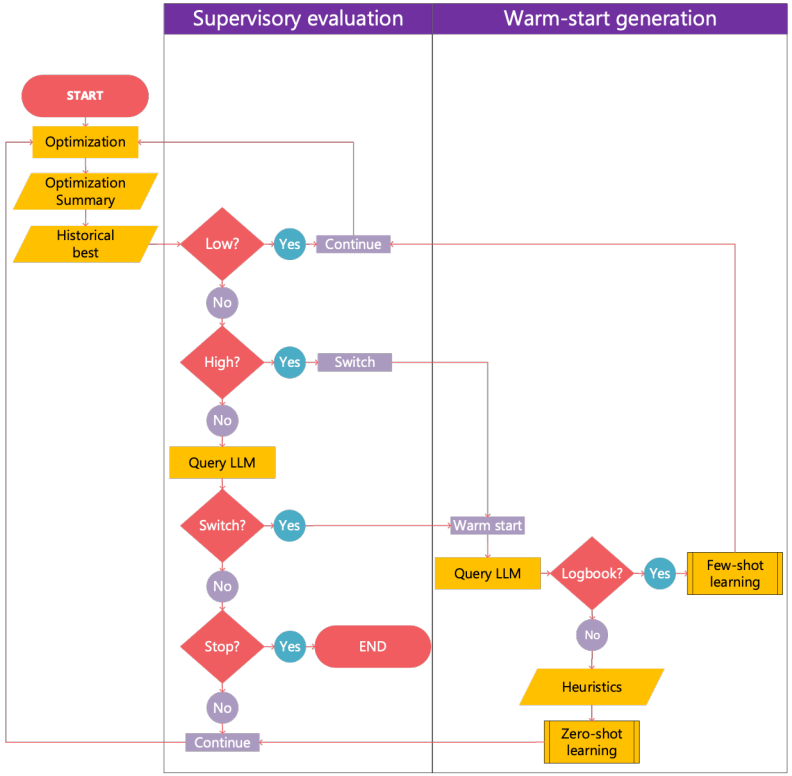
Practical Considerations / Tradeoffs

- **Local compute resources required.** Sufficient CPU/GPU memory is needed to achieve interactive inference latency, although modern workstation hardware is generally adequate for models in the 7B parameter range.
- **Model size limitations.** Locally deployable models are typically smaller than state-of-the-art cloud-hosted LLMs, which may limit reasoning capability for highly complex tasks.
- **Local model maintenance.** Model versions, quantization formats, and inference libraries must be managed by the facility rather than by a cloud provider.
- **Hardware-dependent performance.** Inference speed depends on the available workstation hardware and accelerator support (e.g., CPU, GPU, Apple Metal).
- **Initial deployment effort.** Setting up local inference requires installation and validation of the runtime environment (e.g., llama.cpp, GGUF models).
- **Model updates require validation.** Upgrading to a newer language model should be accompanied by verification to ensure consistent supervisory behavior.
- **Future scalability.** Larger or multimodal models may eventually require more capable hardware if additional reasoning tasks are introduced.

ORION: Orchestrated Real-Time Intelligent Optimization

LLM-enhanced Orchestration Layer:

"Limit LLM to high-level supervisory reasoning and warm-start generation"

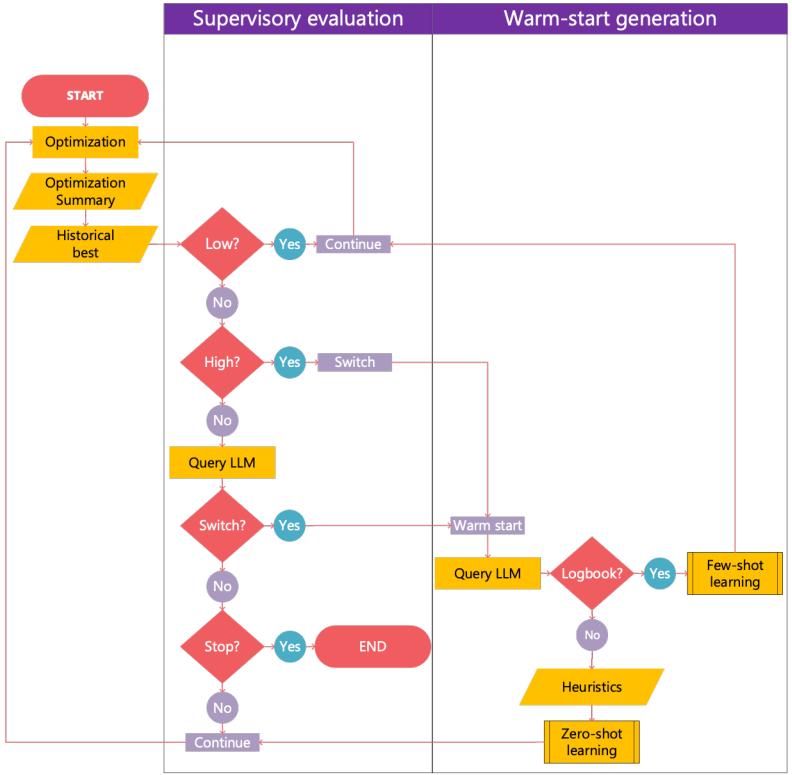


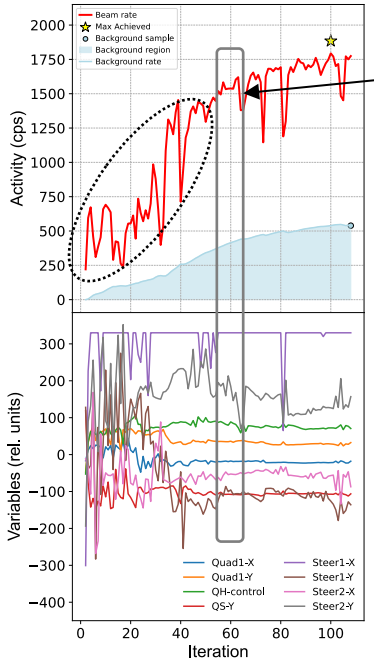
Argonne NATIONAL LABORATORY

ORION: Orchestrated Real-Time Intelligent Optimization

LLM-enhanced Orchestration Layer:

"Limit LLM to supervisory reasoning and warm-start generation"





Start from here

- Accelerate converge
- Reduce time for background buildup. Remember we are working with radioactive beams.

ORION: Orchestrated Real-Time Intelligent Optimization

LLM-enhanced Orchestration Layer:

"By limiting LLM to high-level supervisory reasoning and keeping deterministic BO calculations, ORION can operate without frontier-scale models"

LLM Deployment

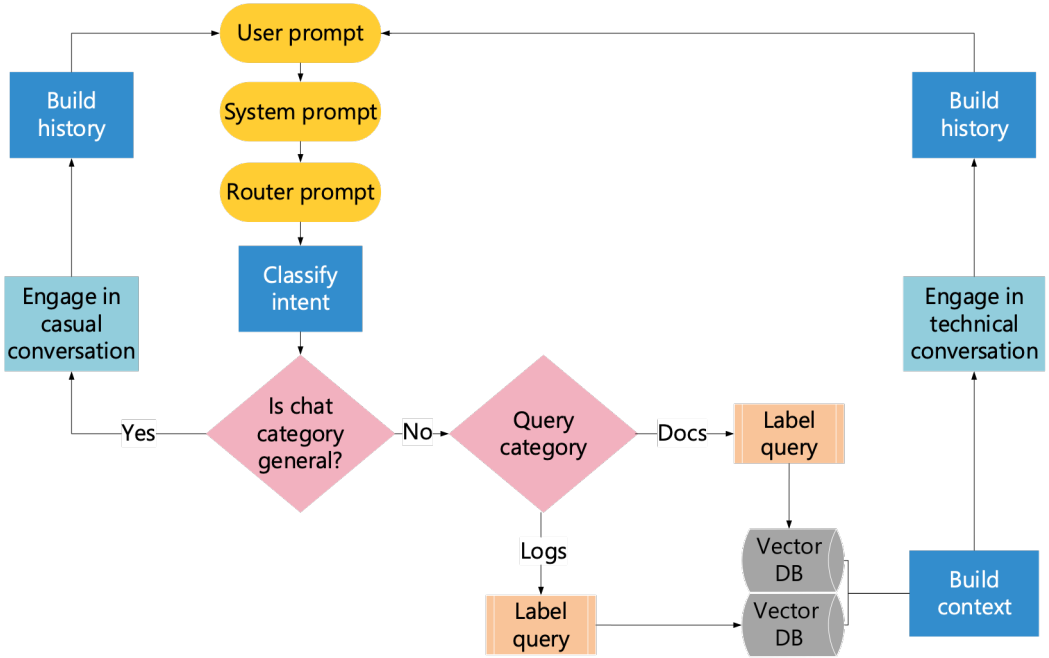
- LLaMA 2 7B model using quantized GGUF wights and run via llama.cpp backend.
- Tested on MacBook Pro with M3 Max chip, 64GB RAM.
- This setup supports inference on a single machine with sufficient memory and compute.
- No need for large GPU cluster.
- Proven performance on live tests that included *three* beamline sections.

Given the limited access to beam time, it would be ideal to have a dedicated facility for further testing and development.

ORION: Orchestrated Real-Time Intelligent Optimization

RAG system:

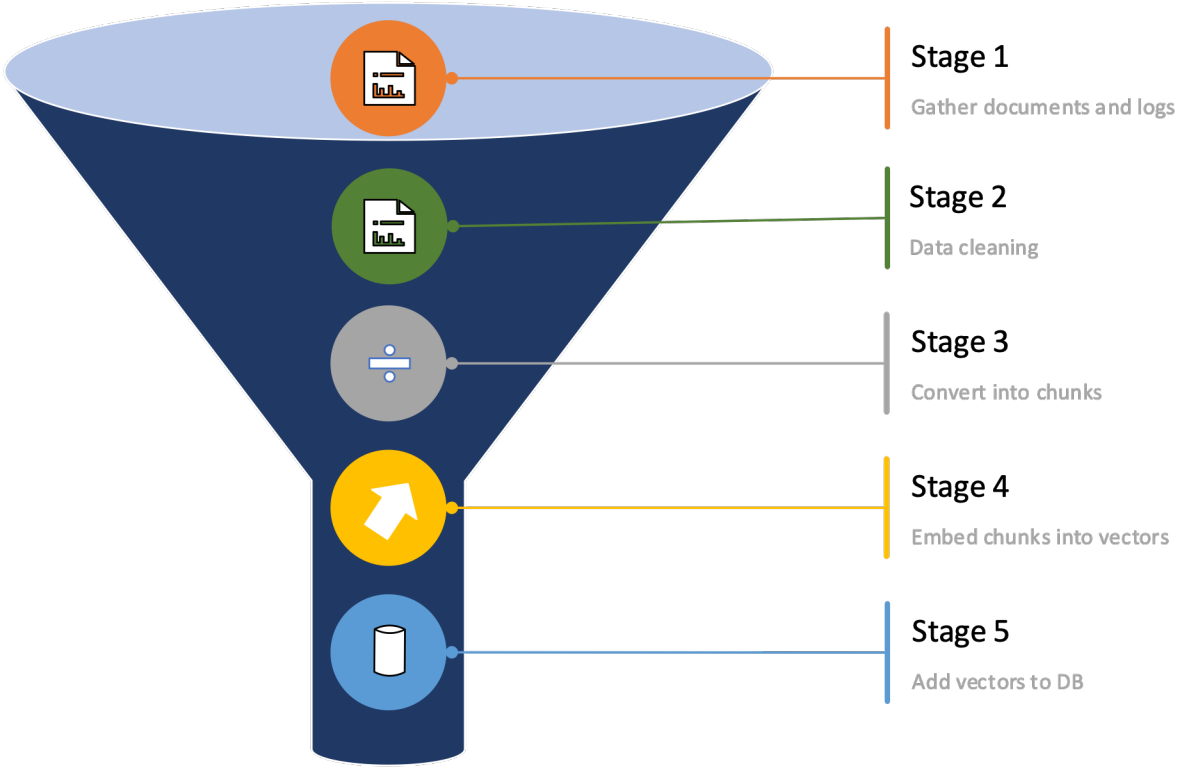
“Independent RAG system to provide contextual information”



Argonne NATIONAL LABORATORY

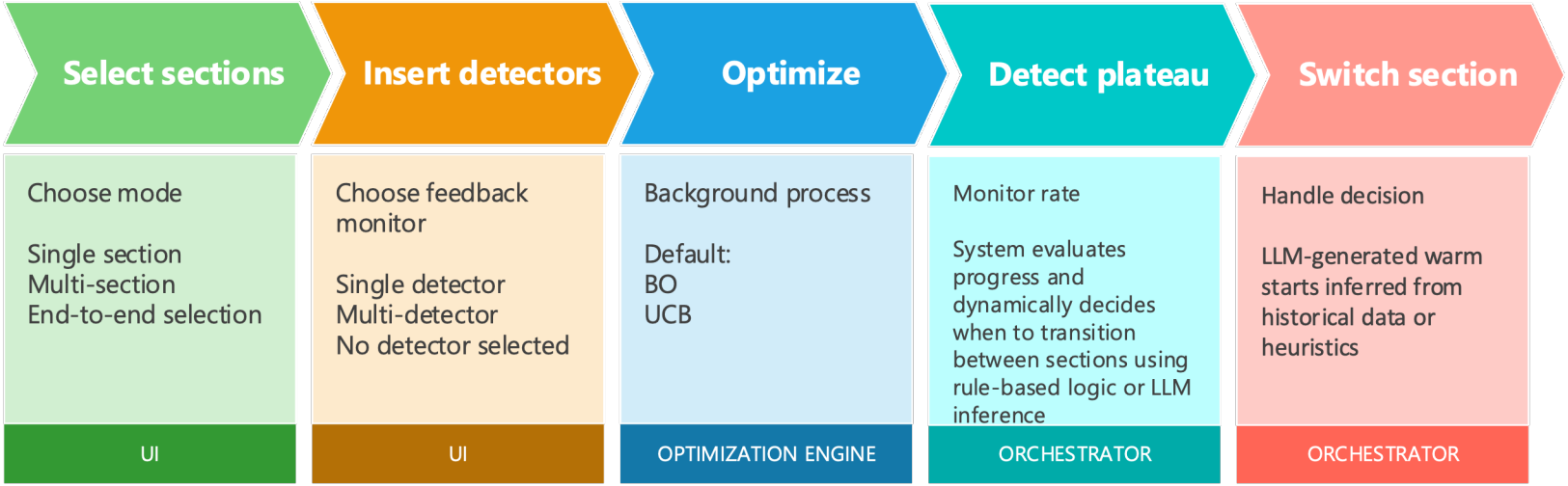
ORION: Orchestrated Real-Time Intelligent Optimization

Vector DB ingestion:



ORION: Orchestrated Real-Time Intelligent Optimization

Workflow:

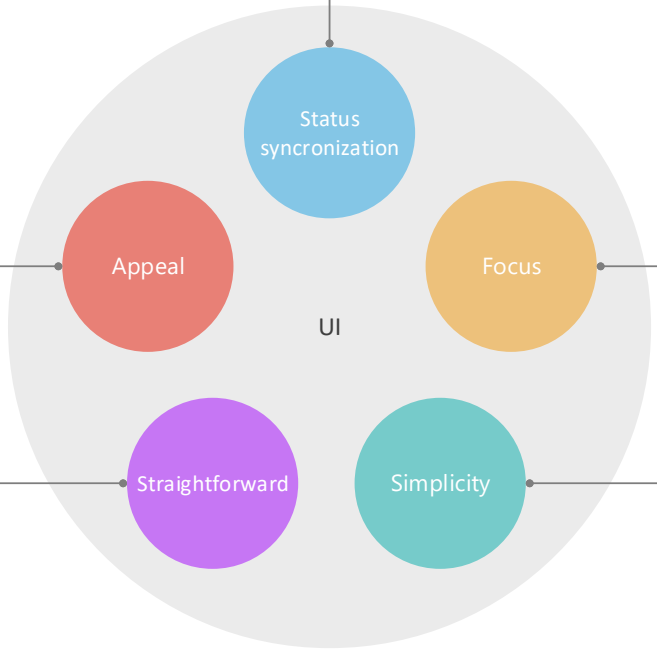


ORION: Orchestrated Real-Time Intelligent Optimization

UI design philosophy:

"Design the UI around the operator (non-expert) experience"

Visuals must reflect the underlying optimization and hardware status.



Operators need an appealing app to improve their experience and encourage long-term use.

Focus on the essentials and leave the rest out. Deliver beam from source to experiment.

Perform task in as few clicks as possible. Make operators lives easier.

Avoid presenting expert-level configuration parameters.

ORION: Orchestrated Real-Time Intelligent Optimization

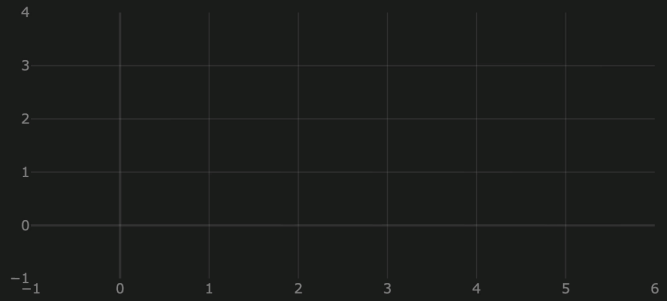
Demo:

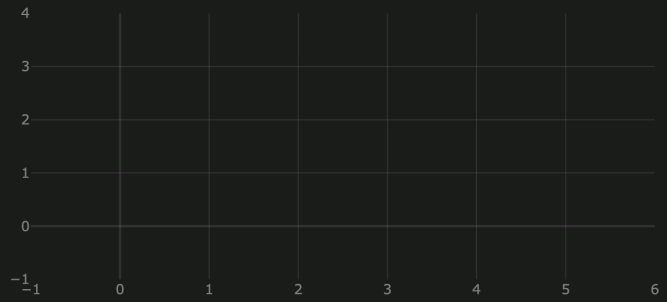
ORION

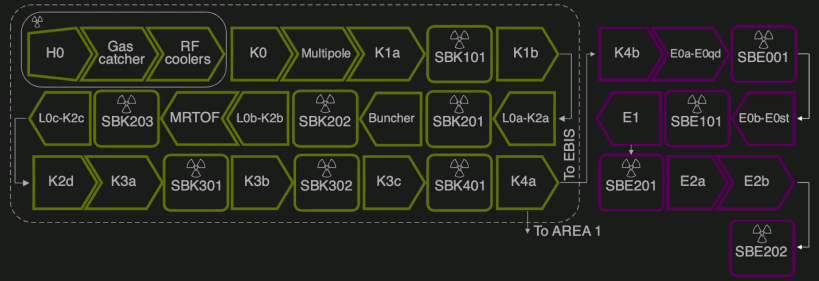
127.0.0.1:8050

ORION - Orchestrated Real-time Intelligent OptimizatioN

START STOP RESET Environment: Select... Generator: Select... Assistant Model: Select...







Hi. I'm ORION, your beamline assistant.

Ask about logs, operations, or manuals...


Argonne
NATIONAL LABORATORY