Brookhaven National Laboratory

U.S. DEPARTMENT of ENERGY

The Brookhaven AI Inference (BRAIN) Platform

Secure, On-Premise AI for Scientific Computing and Operations

Meifeng Lin

Computational Science Department, CDS

SCDF Stakeholders Meeting

08/13/2026

@BrookhavenLab

Vision and Objectives

Primary Goal: Deploy secure, locally hosted Large Language Models (LLMs) to support Brookhaven's research and administrative workflows without compromising data privacy.

Key Drivers:

- **Data Security:** Keep sensitive research data strictly on-premise.
- **Customizability:** Fine-tune models for specific scientific domains and augment retrieval using internal BNL documentation (e.g. SBMS).
- **Accessibility:** Lower the barrier to entry for AI integration across various lab departments and scientific projects.
- **Cost:** Rising costs with commercial AI models and platforms.

Hardware Infrastructure Available

Current Compute Resources:

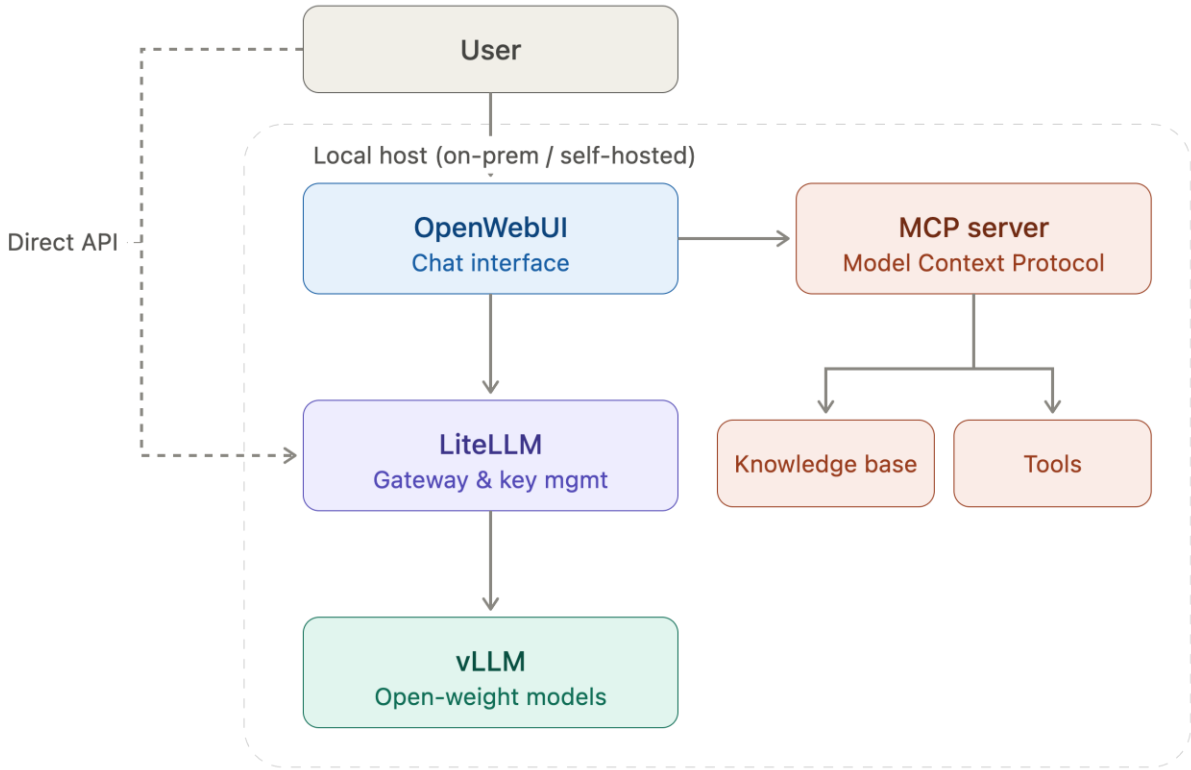
- GPUs: SciServer (**40** H100 GPUs) + new NVIDIA B300 GPU server
- Storage: PBs of Luster storage
- Networking: Leveraging BNL's high-bandwidth internal network
- Some reserved for HPC workflows; some repurposed for AI inference.

Future Scaling: Planned expansion of the GPU footprint to support larger, more complex models.

- Potential to install more GPU racks in the data center in the future.

Services Provided

1. **Internal Chat Interface:** A secure, web-based UI for general inquiries, drafting emails, and coding assistance. Inclusion of image generation capabilities through Flux.1.
2. **API Access:** RESTful APIs for researchers to integrate LLM capabilities directly into their scientific workflows or AI agents.
3. **Document Retrieval (RAG):** "Chat with your data" capabilities to securely search and summarize large volumes of internal PDFs, reports, and technical manuals.
4. **MCP Server**: Hosting internal knowledge base and specialized tools



BRAIN Architecture

Models Currently Hosted (US Models Only)

Model	Provider	Practical best uses	Context window / prompt limit	Modalities	Reasoning / generation features
Llama 4 Scout 17B-16E	**Meta**	Extreme long-context analysis, huge document collections, multimodal RAG, repository-scale code analysis	**Up to 10M tokens (limited to 32K tokens in current deployment)**	Text + image → text/code	General reasoning, coding, image understanding, multimodal grounding; exceptional long-context retrieval
Gemma 4 26B-A4B	**Google**	Efficient general-purpose assistant, coding, scientific assistance, multimodal RAG, moderate-scale agents	**Up to 256K tokens (limited to 128K token in current deployment)**	Text + image → text/code	Configurable thinking, native function calling, coding, reasoning, system-role support
GPT-OSS-120B	**OpenAI**	Deep reasoning, scientific problem solving, difficult coding, structured outputs, tool-driven agents	**Up to 131,072 tokens**	Text → text/code	Configurable reasoning effort, native function calling, tool use, structured output, fine-tuning
Nemotron 3 Super 120B-A12B	**NVIDIA**	Production agentic AI, multi-agent systems, coding, RAG, tool orchestration, high-throughput reasoning	**Up to 1M tokens**	Text → text/code	Reasoning-budget control, tool calling, planning, long-running and collaborative-agent optimization
Nemotron 3 Ultra 550B-A55B (Preview)	**NVIDIA**	**Highest-accuracy reasoning**, complex scientific/code reasoning, mission-critical multi-step agents, deep planning, high-stakes RAG, very long-context analysis	**Up to 1M tokens (limited to 131,072 tokens in current deployment)**	Text → text/code	Configurable reasoning mode/budget, tool use, complex agentic workflows; native speculative decoding via MTP; trained with SFT, RL and multi-teacher on-policy distillation
FLUX.1 image-generation	**Black Forest Labs**	Text-to-image generation, scientific illustrations, presentation graphics, outreach, UI assets, concept art	**Up to 512 input tokens**	Text → image	Strong prompt adherence, typography, style diversity, variable resolutions/aspect ratios.

Brookhaven National Laboratory

Modified from a summary generated with ChatGPT 5.6

How Good Are These Models?

LEADERBOARD — RANKED BY AVERAGE SCORE

RANK	PROVIDER	AVG SCORE	PASS RATE	AVG LATENCY	TOTAL COST	TOTAL TOKENS	BENCHMARKS	N
1	**nemotron-3-ultra-550b-a55b** openai:chat	0.941	93.7%	42724ms	$0.0000	3,954,663	8/8	4246
2	**claude-sonnet-4-6** anthropic:messages	0.935	92.5%	9112ms	$6.1223	819,300	8/8	4246
3	**gemma-4-26b** openai:chat	0.917	90.9%	1586ms	$0.0000	734,452	8/8	4246
4	**gpt-oss-120b** openai:chat	0.913	90.2%	3235ms	$0.0000	1,651,568	8/8	4246
5	**nemotron-3-super-120b** openai:chat	0.887	88.1%	9501ms	$0.0000	2,791,095	8/8	4246
6	**llama-4-scout-17b** openai:chat	0.878	86.6%	3344ms	$0.0000	687,270	8/8	4246
7	**gpt-5.4-nano** azure:chat	0.839	83.2%	5107ms	$0.7432	987,596	8/8	4246

Benchmark	Capability Tested	Format	Questions	Status
OpenBookQA	Multi-step science reasoning + common sense	4-way MCQ	500	Complete
SciQ	Physics, chemistry, biology knowledge	4-way MCQ	1,000	Complete
CoLA	Linguistic acceptability / grammar	Binary MCQ	500	Complete
MMLU	Broad knowledge across 57 subjects	4-way MCQ	500	Complete
RACE	Reading comprehension	4-way MCQ	500	Complete
HumanEval	Python code correctness (execution-graded)	Pass/fail	164	Complete
IFEval	Instruction-following compliance (deterministic)	Verifiable checks	541	Complete

Detailed Scores

SCORE MATRIX — PROVIDER × RUN

PROVIDER	BROAD GENERAL KNOWLEDGE (STEM)	GRADE SCHOOL SCIENCE QUESTIONS	INSTRUCTION FOLLOWING COMPLIANCE (IFEVAL) (LOOSE)	INSTRUCTION FOLLOWING COMPLIANCE (IFEVAL) (STRICT)	LINGUISTIC ACCEPTABILITY JUDGMENTS	PYTHON FUNCTION CORRECTNESS (HUMANEVAL)	READING AND COMPREHENSION	SCIENCE EXAM QUESTIONS	OVERALL
nemotron-3-ultra-550b-a55b	0.93 93% pass	0.96 96% pass	0.96 95% pass	0.96 94% pass	0.85 85% pass	0.98 98% pass	0.91 91% pass	0.97 97% pass	0.94 94% pass
claude-sonnet-4-6	0.90 90% pass	0.95 95% pass	0.95 93% pass	0.92 87% pass	0.86 86% pass	0.98 98% pass	0.93 93% pass	0.98 98% pass	0.93 93% pass
gemma-4-26b	0.84 84% pass	0.94 94% pass	0.94 91% pass	0.93 89% pass	0.85 85% pass	0.97 97% pass	0.91 91% pass	0.96 96% pass	0.92 91% pass
gpt-oss-120b	0.90 90% pass	0.95 95% pass	0.92 88% pass	0.90 85% pass	0.83 83% pass	0.98 98% pass	0.86 86% pass	0.97 97% pass	0.91 90% pass
nemotron-3-super-120b	0.82 82% pass	0.95 95% pass	0.91 89% pass	0.89 86% pass	0.81 81% pass	0.88 88% pass	0.87 87% pass	0.97 97% pass	0.89 88% pass
llama-4-scout-17b	0.79 79% pass	0.93 93% pass	0.92 87% pass	0.90 84% pass	0.80 80% pass	0.83 83% pass	0.90 90% pass	0.96 96% pass	0.88 87% pass
gpt-5.4-nano	0.86 86% pass	0.94 94% pass	0.70 68% pass	0.69 66% pass	0.82 82% pass	0.88 88% pass	0.86 86% pass	0.96 96% pass	0.84 83% pass

Top Model Performance Summary

- **Leader: nemotron-3-ultra-550b** (0.94 overall) — Exceptional instruction following and science proficiency.
- **Top-Tier: gemma-4-26b** (0.92) — Highly balanced performance across reasoning and science.
- **Key Strengths:** Universal high performance in **Python (HumanEval)** and **Science (SciQ)** across all top models.
- **Primary Weakness:** All top models show a relative dip in **Linguistic Acceptability (CoLA)**.

Rollout Schedule & Timeline

Phase 1: Proof of Concept and Beta Launch (Q4 FY2026)

- Deploy initial hardware infrastructure.
- Test Chat UI and API access with selected groups of users (scientists and ops staff).
- Gather feedback on latency, model accuracy, features and workflow integration.

Phase 2: General Availability (Q1 FY2027)

- Lab-wide rollout of LLM services, utilizing new hardware procured.
- Introduce domain-specific fine-tuning options for individual departments.

Phase 3: Ongoing Optimization and Scale-up (Q2 FY2027 & Beyond)

- Scale hardware, update to newer model architectures, and expand integration with broader BNL resources.

Current Status

Service	Status	Details	Access/Authentication
Test Chat UI	✅ Beta Testing	inference1.sdcc.bnl.gov — all 6 models available (incl. bnl-nemotron-3)	BNL Domain Account + VPN or wired network only
REST API (LiteLLM)	✅ Beta Testing	Base: https://inference0-api.sdcc.bnl.gov (OpenAI compatible endpoints)	Base URL + API Key + VPN or wired network only
API Key Management	✅ Beta Testing	Generate/revoke keys at https://inference0-api.sdcc.bnl.gov/km **30-day lifetime**	BNL Domain Account + VPN or wired network only
BNL Knowledge Base Integration	✅ Beta Testing	1. Use the **bnl-nemotron-3** model— direct URL refs (Recommended) 2. **BNL-KB** tool — enable in Tools menu with any model (no URL refs)	Self-service via chat UI
MCP Server	⚠️ Limited Beta Testing	Model Context Protocol — internal knowledge bases, specialized tools	Select groups
BRAIN Portal	🔒 In Development	brain.bnl.gov — user guide, leaderboard, FAQs (not yet active)	BNL Domain Account + VPN or wired network only

🔒 **Important Caveats:** Test instance data may not migrate to the production BRAIN instance.

Acknowledgments

Core Technical Team:

- Mohammad Atif – MCP Server, BNL knowledge integration
- Costin Caramarcu – Initial prototypes of Open WebUI, vLLM/LiteLLM
- Kriti Chopra – Model performance benchmarking
- Zhihua Dong – Deployment, hardware infrastructure
- Mikhail Titov – Web pages, user guide, GitHub
- Tianle Wang – Computational performance benchmarking, optimization

Advisory:

- Nick D'Imperio
- Tony Wong
- Shinjae Yoo

We thank ITD for their prompt support of the changing service requirements!