



BernieAI Pilot Findings

Yousseff Abed,
Cebastian Santiago Hernandez
Larence Livermore National Labs

PRESENTED BY



SPONSORS



Stantec



TETRA TECH

Jacobs



POND

BernieAI: 15-Month Pilot Report

National Nuclear Security Administration (NNSA)
Yousseff Abed, Lawrence Livermore National Laboratory (LLNL)
BernieAI Lead and BUILDER Project Manager

Stephanie Greco, Lawrence Livermore National Laboratory (LLNL)
Deputy Program Leader for IAP

Cebastian Santiago Hernandez, Lawrence Livermore National Laboratory (LLNL)
BernieAI Lead Developer

Mitchell Haraburda, NNSA HQ Operations & Maintenance Division (NA-914)
BernieAI Program Lead

BernieAI Pilot:

Transforming NNSA infrastructure decision-making

BernieAI was a 15-month NNSA pilot that began in CY2023 to evaluate AI solutions for strategic investment decisions often hampered by fragmented data, aging assets, and reliance on anecdotal evidence over data-driven analysis.

Purpose

Evaluate if AI could enhance infrastructure planning and investment across the NNSA enterprise

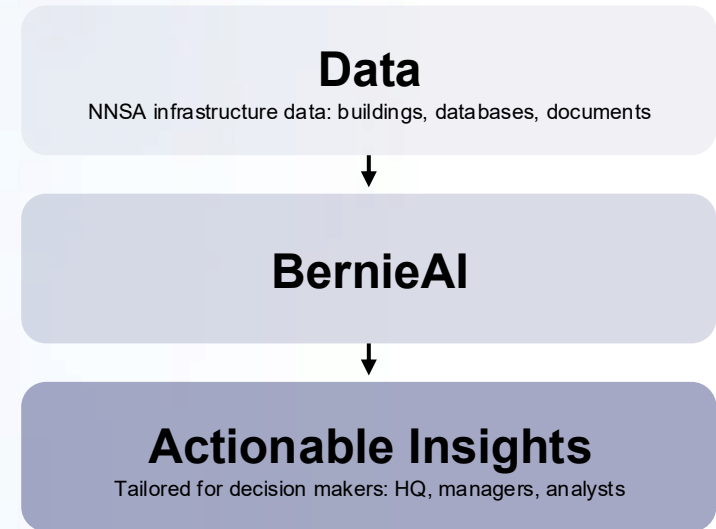
Build a prototype of BernieAI to connect enterprise data with advanced AI for better decision support

Approach

Collaborative Multi-disciplinary, cross-site teams (LLNL, Y-12, NNSA HQ).

Focus on secure, transparent, and reproducible AI and avoiding “black box” solutions

Develop a web portal powered by a multi-agent AI system, connected to enterprise data



BernieAI demonstrated a secure, transparent, and collaborative AI system that could unlock the full value of NNSA’s infrastructure data, **enabling smarter, faster, and more defensible investment decisions.**

LLMs & Frameworks: The brains behind BernieAI

Bernie is built on a hybrid-model strategy, combining open-source and proprietary LLMs for flexibility, security, and access to cutting-edge capabilities.

Small Language Model (SLM)

Dec 2023

Pre-pilot
llama.cpp
on a laptop

Feb - Jul 2024

EC2 Linux Server
on AWS GovCloud

Aug - Sept 2024

Mistral on EC2
7 Billion Parameter

Discussion on transition
to LLM and further SLM
investigation

Oct 24 - Jan 2025

Hermes on EC2
7 Billion Parameter

Large Language Model (LLM)

Feb - Apr 2025

Llama 3.3 70 billion
parameter supported
by AWS on P4
instance

May - Jul 2025

GPT 4.0 on LivAI
Microsoft Azure API
>1 trillion parameter

Aug 25 - Present

GPT 4.1 on LivAI
Microsoft Azure API
>1 trillion parameter

Critical Considerations for NNSA Enterprise Model Selections



Context Window Size



Model Collapse



Privacy + Security



Quality Assurance

Bernie's agentic, hybrid LLM architecture, supported by scalable cloud hardware ensures secure, high-performance, and future-proof AI capabilities for NNSA infrastructure operations.

Hardware Evolution: From laptops to scalable cloud solutions

Windows Laptop December 2023-January 2024

- + Initial open-source LLMs restricted to local machines
- Lack of GPU hardware = slow performance, bottlenecks

Linux Server January 2024 – May 2024

- + Infrastructure shift was needed, Bernie moved to “Poffin”
- Still CPU-only, robust models require GPU acceleration

EC2 Server + GovCloud Summer 2024 - Current

- + Support for larger models and improved performance
- Operational expenses increased with model scale

Infrastructure Choices & Rationale

AWS GovCloud was ultimately chosen for secure, scalable hosting behind the LLNL firewall but in the future, docker containerization enables portability across cloud providers and potential on-premises deployment opportunities.

EC2 was chosen over managed services like SageMaker for greater control, transparency, and learning opportunities. On-premises supercomputing (Livermore Computing) was also considered for future training/fine-tuning, but not viable for persistent hosting due to environment constraints.

Evaluation of hardware options to optimize for both performance and budget are ongoing

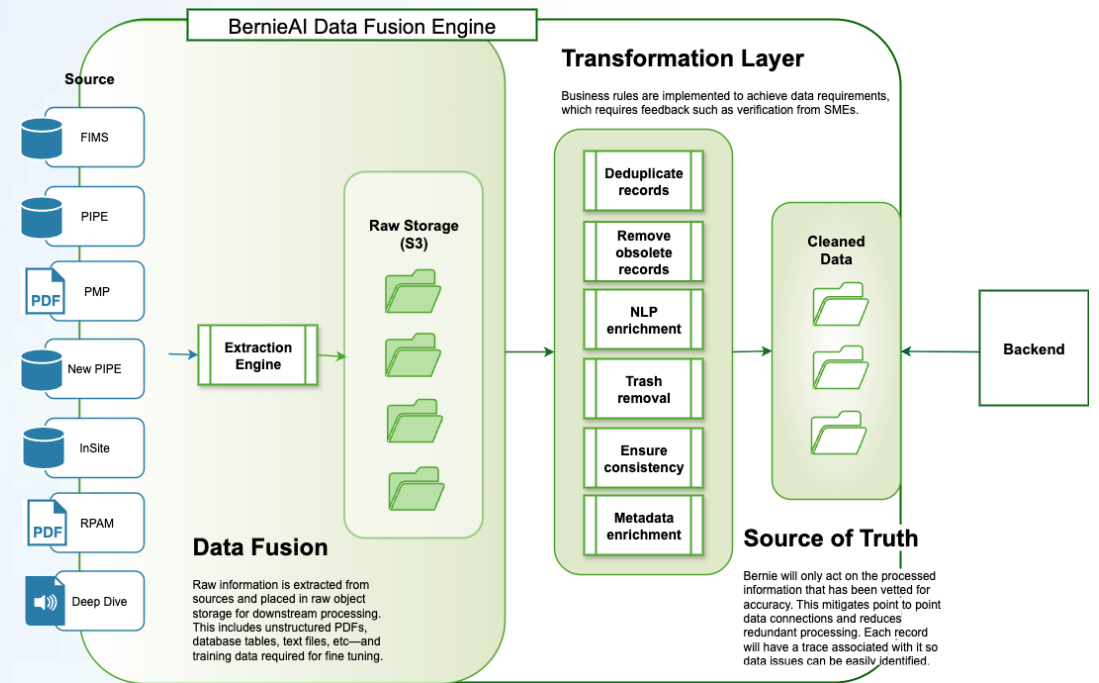
Migrating BernieAI to scalable, GPU-enabled cloud infrastructure was essential for supporting advanced AI models and enterprise growth, but requires continuous balancing of performance, cost, and flexibility.

Data Strategy: Integration, quality, and scalability

BernieAI will utilize a Data Fusion Engine (DFE) strategy: a centralized, scalable, efficient data platform designed to improve the quality, accessibility, and governance of data for BernieAI and NNSA enterprise analytics.

DFE Benefits Include:

- Provides a single, auditable “source of truth” for all data consumed by BernieAI.
- Improves performance and scalability by separating transactional and analytical workloads.
- Ensures data quality, consistency, and observability across the entire data lifecycle



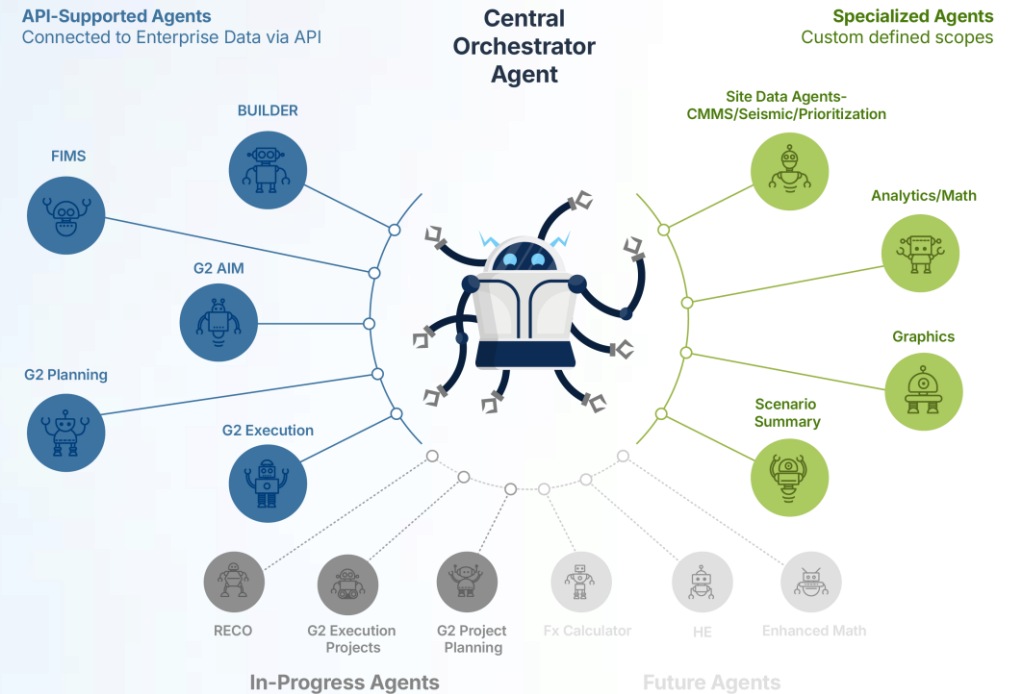
Bernie’s data strategy unifies structured and unstructured sources through rigorous integration, validation, and scalable architecture, laying a foundation for trustworthy, enterprise-wide AI analytics and decision-making.

LLMs & Frameworks: Multi-agent framework brings the experts to you

Agent-based design transformed BernieAI from a static chatbot into a dynamic, collaborative problem-solver that mirror real-world analytical teamwork.

Bernie's multi-agent framework boosts efficiency and scalability by assigning **Agents**, like specialized employees, to each task.

An **Orchestrator Agent** acts as the liaison, ensuring the right experts are called in to deliver the best answers.



The suite of current and future Agents are foundational to the success of future Expert Agents

BernieAI's data strategy unifies structured and unstructured sources through rigorous integration, validation, and scalable architecture, laying a foundation for trustworthy, enterprise-wide AI analytics and decision-making.

Retrieval Augmented Generation (RAG): Enhancing AI with real world data

Bernie's RAG system is a vector database and agent that processes user queries, retrieves relevant documents, and fuses information for accurate, context-rich responses.

Hybrid Retrieval Yields Best Results

Combining semantic vector search with keyword matching, using LanceDB ensured both meaning and specificity in retrieved information

Modular Designs Boosts Performance

Processing sub-queries in parallel and asynchronously reduced response time and allowed for scalable, efficient retrieval from large data sets

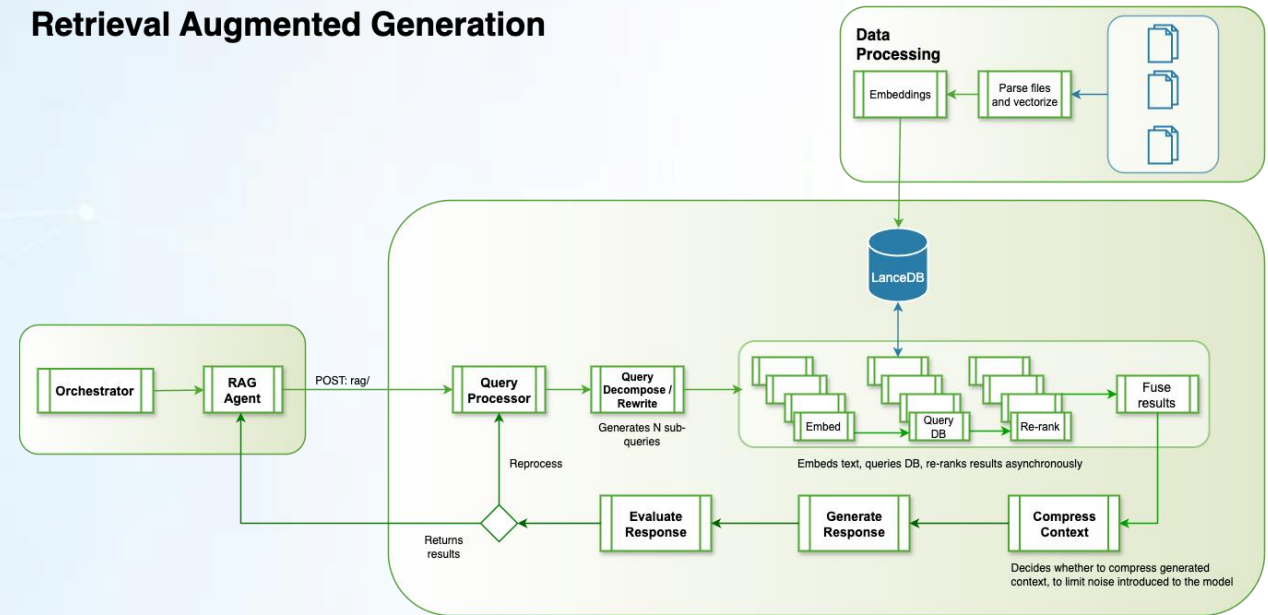
Document Chunking & Pre-Processing Matter

Sophisticated chunking with Docling helped preserve context and meaning for complex documents with tables, diagrams, and images

Continuous Tuning is Needed

Early pure retrieval approaches had limitations so iterative improvements and SME feedback were key to refining the RAG pipeline

Retrieval Augmented Generation



RAG empowers BernieAI to deliver accurate, trustworthy answers by combining advanced language models with real-time retrieval from unstructured authoritative enterprise data, ensuring outputs are grounded in the most relevant and current information.

Portal Development: Making complex AI workflows usable

The BernieAI portal succeeded by prioritizing intuitive design, maintaining transparency and security, and continuously assessing user feedback to deliver a platform that builds trust and meets both enterprise and site user needs.

Current Portal Tech Stack:

Java backend, Angular Material frontend, REST APIs to agentic system

Key Features:

- Conversational interface with agent dropdown (e.g., general, Y-12, doc agent).
- Getting Started tab, Dev Sandbox, and Wiki/About pages for onboarding.
- Context toggle to show data sources, agents used, and per-request cost.
- Speech-to-text, stop response, print to PDF, light/dark theme.
- Chat history (v1) with roadmap toward searchable, bookmarkable knowledge.

The screenshot displays the BernieAI portal interface. At the top, there's a navigation bar with options like 'Confluence', 'Spaces', 'People', 'Questions', 'Calendars', 'Analytics', and 'Create'. Below this, the main content area is titled 'Bernie-AI' and includes a 'Project Overview' section with a mission statement, vision, approach, and details. A 'Project Team' list is also visible. On the left, there's a 'Getting Started' section with a 'Hello! I'm Bernie-AI. Use the buttons below to learn how I can help you.' message. Below this, there are buttons for 'Bernie-AI overview & capabilities', 'Data sources', 'Chart generation', 'Writing effective prompts', 'Basic prompt structure', 'AI agents overview', and 'Recursive prompting overview'. The bottom part of the screenshot shows a chat window with a message input field, a dropdown menu for selecting an agent (currently 'Bernie-AI'), and a 'Voice Recognition' button.

A successful AI portal is more than a chat window—it's an evolving, transparent, and secure workspace shaped by continuous user engagement, balancing technical power with intuitive, trustworthy design

User Focus Groups: Co-designing with our users

Subject matter experts such as the LLNL R&D Modeling Group and Y-12 Plant Health were brought in early and were critical to shaping realistic use cases and prompts, User and Subject Matter Expert collaboration was not just a phase, but a continuous, iterative process that shape important elements of Bernie from prompt engineering to data validation and feature prioritization.



R&D Modeling Group

- Helped reframe demos from “tech showcase” to “analyst use case.”
- Developed lifecycle “Train of Thought” prompts (e.g., “When should we replace this facility?”).
- Drove narrative creation and V&V for metrics like RPV, DM, BCI, MDI.



Y-12 Deep Dive

- Joint team linked FLOC across BUILDER, G2, Plant Health, historic maintenance.
- Co-developed Train of Thought for “How best to invest 20M dollars in fabrication fans”
- Successfully demonstrated BernieAI live in MAP Deep Dive, with strong stakeholder enthusiasm.



Verification and Validation Ensures Trust

Literal and contextual V&V processes, led by SMEs, helped ensure data accuracy, reduce hallucinations, and clarify ambiguous term while narratives and context documentation improved the AI’s ability to interpret complex, nuanced data.

Active, ongoing engagement with users and SMEs through demos, collaborative prompt design, and feedback was essential for improving AI performance, and ensuring the system met real analyst needs.

Cybersecurity: Security First, Safeguarding Data & Access

Bernie protects confidentiality by encrypting all data during storage and transmission, and by logging system activity. Staff monitors the logs to quickly detect and respond to potential security threats.



Enterprise-Grade Security Standards

- BernieAI complies with DOE, FISMA, Privacy Act requirements and follows strict cybersecurity orders for data protection.
- Multi-tenant infrastructure enforces data segregation and privacy across institutional boundaries.



Layered Approach to Access Control

- Federated authentication (OneID SSO), multi-factor authentication, and role-based access control (RBAC) restrict access to authorized users only.
- Permissions, user roles, and data partitions are tightly managed to ensure least-privilege access.



Red Teaming and Penetration Testing

- Proactive security assessments identified and mitigated vulnerabilities (e.g., prompt injection, broken access control).
- Regular penetration testing and prompt implementation of recommendations improved resilience.



Adaptive Security Posture

- Security protocols evolved with project needs, including bifurcated data for site-specific agents and question redirects for unauthorized queries.
- Ongoing audits and privilege reviews support a zero-trust model and future scalability.



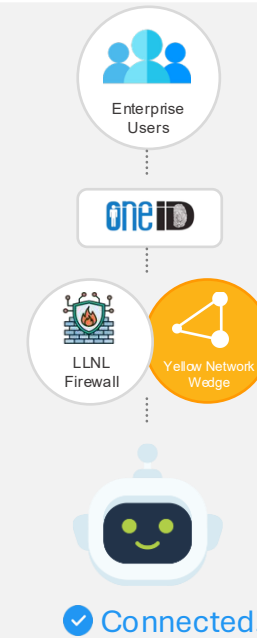
Network and Data Classification

- Hosting and access decisions guided by CUI data classification and network policies (yellow network, VPN solutions).
- Collaboration with external sites required creative solutions to meet security and compliance requirements.



Fine Tuning and Model Collapse Risks

- Fine-tuning is performed only on vetted, non-sensitive data to prevent leakage of confidential information.
- Continuous monitoring and governance mitigate risks like model collapse and data contamination.



The **“Yellow Network Wedge”** is an external-facing network behind the LLNL firewall enabling enterprise contacts to use OneID to access and use BernieAI

BernieAI’s robust, multi-layered data security framework ensures secure access, data integrity, and compliance supporting trusted AI adoption across the NNSA enterprise.

Costs and Sustainability: Managing Resources for BernieAI

Cloud GPU costs far exceeded estimates:

AWS GPU hosting drove hardware expenses up by ~1,300%, requiring manual shutdowns to control spend.

Energy tracking is critical

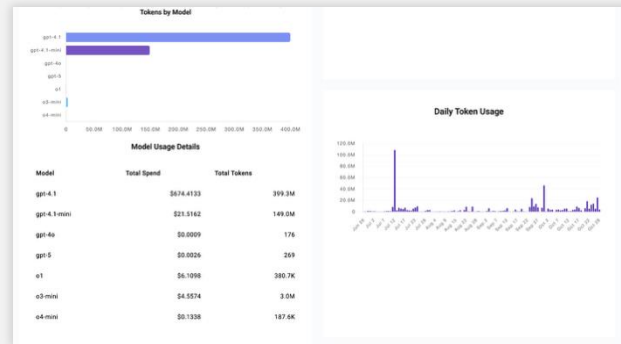
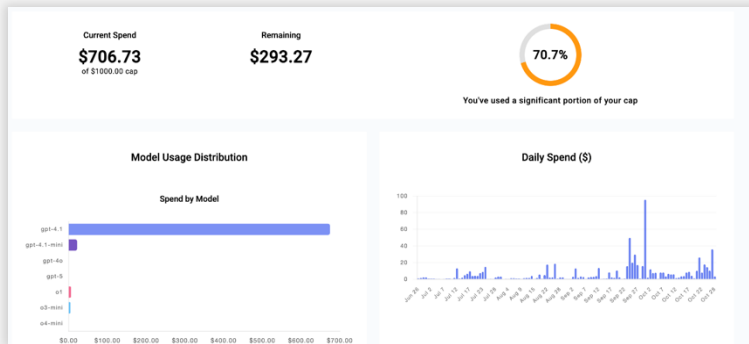
With direct measurement unavailable, the team used model-based estimates to monitor energy use per query.

Switch to managed API reduced complexity:

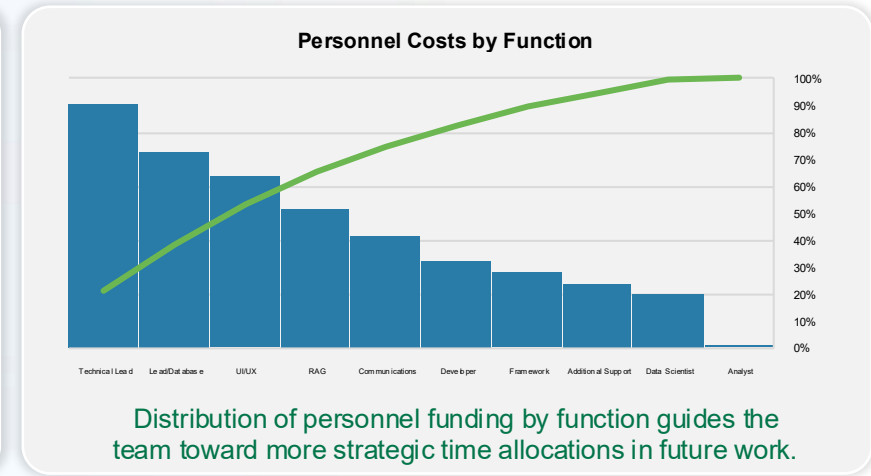
Moving to LivAI GPT-4.x shifted costs from hardware to per-token usage, improving reliability but adding pricing uncertainty.

Dedicated and well trained staff is required

Pilot team has the necessary experience and requisite skills to support an enterprise-wide AI initiative like BernieAI



The transition to the LivAI Microsoft Azure API instance of GPT4.0 substantially reduced the compute costs by shifting the primary expense from hardware (right) to token-based usage (left).



Distribution of personnel funding by function guides the team toward more strategic time allocations in future work.

Bernie demonstrated clear mission value, but exposed GPU & cloud choices can drive costs up by an order of magnitude, making long-term sustainability dependent on disciplined infrastructure strategy, usage controls, and dedicated staffing.

Putting it All Together: Analysis and Key Findings



Data Quality is Critical

Clean, well-governed data is the foundation for reliable AI outcomes.



Live Data Integration

Direct connections to enterprise data sources minimize hallucinations and improve accuracy.



Expertise Matters

Progress depends on deep developer skills and active SME involvement for V&V



Model Strategy

Flexible “loose affiliation” with models is essential. Frontier models like GPT-4.x are preferred as open-source options currently lack HQ approval.



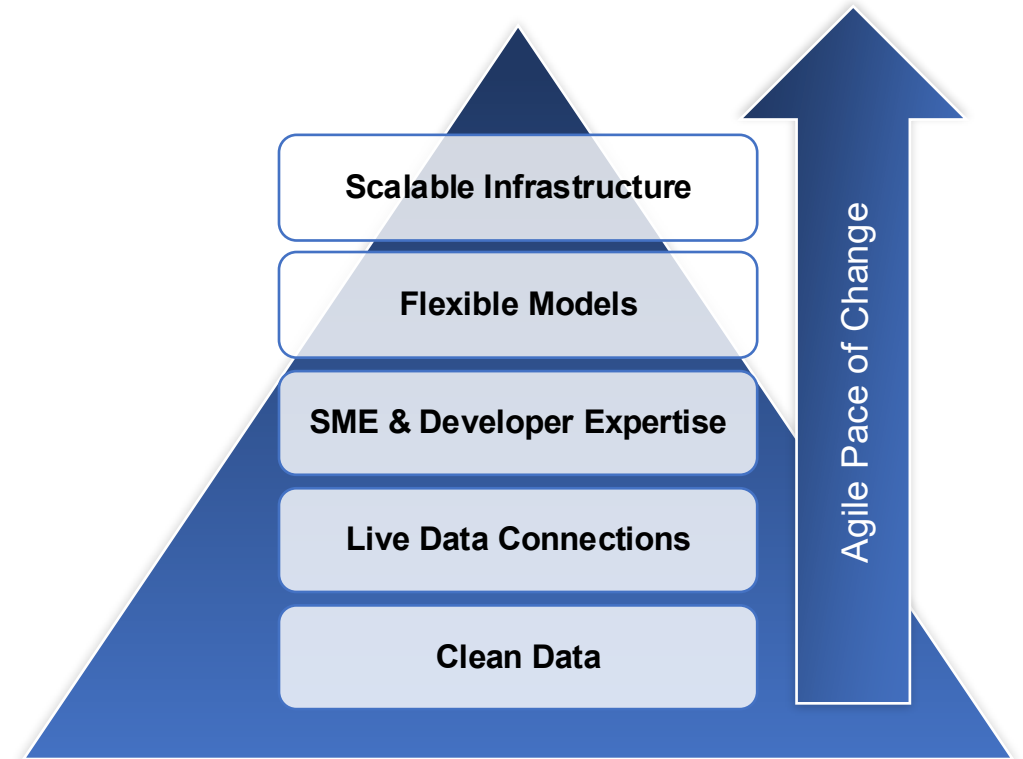
Cloud GPU Challenges

AWS P4 instances are costly and scarce, creating delays and overruns.



Rapid Evolution

The AI landscape is constantly evolving. Teams must focus on continuous learning and fast iteration to keep up with industry.



Ongoing success with BernieAI requires scalable infrastructure, flexible model choices, strong technical/SME teams, live-data access, and continuously improved data quality.