

AWS's AI Bet

What re:Invent 2025 Actually Means

Adam Gordon Bell · Community Engineer ·
Pulumi.com

re:Invent 2025: The Firehose

- Nova Lite, Pro, Sonic, Omni
- Nova Forge
- Trainium3 UltraServers
- AgentCore updates
- Nova Act

The Pattern



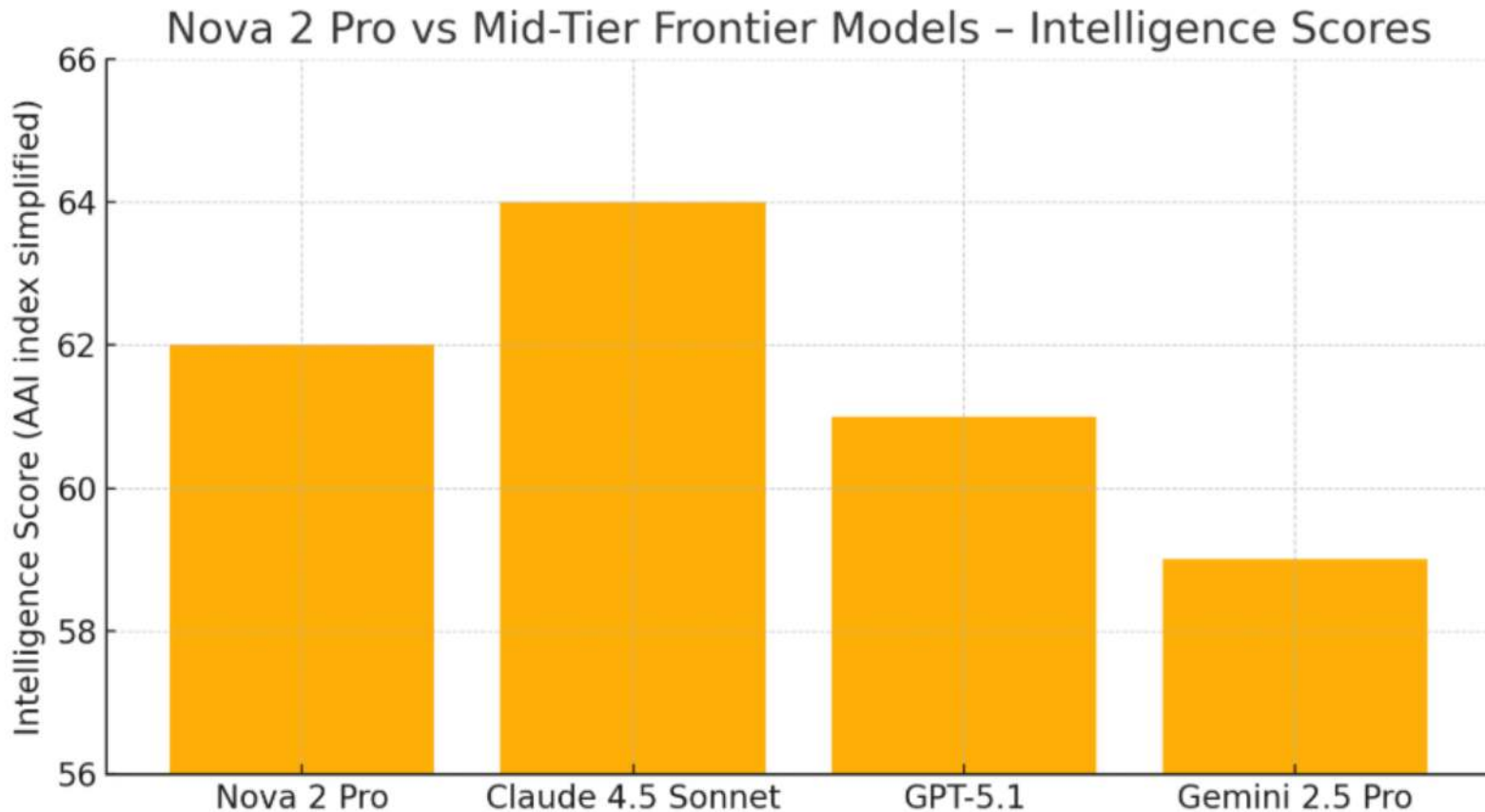
The Thesis

- AWS built a vertically integrated agent-training pipeline.
- It's expensive, ambitious, and not for everyone.

The Models: Nova 2

- Nova 2 Lite, Pro, Omni, Sonic
- Competitive with Claude, GPT, Gemini
- But AWS isn't leading with benchmark wars
- The models aren't the story. The system is.

Nova 2 Benchmarks



Rent the Lab: Nova Forge



Nova Forge: What You Get

- Continued pretraining on your data
- Instruction tuning (SFT)
- RL-style preference optimization
- AWS handles cluster management and hosting

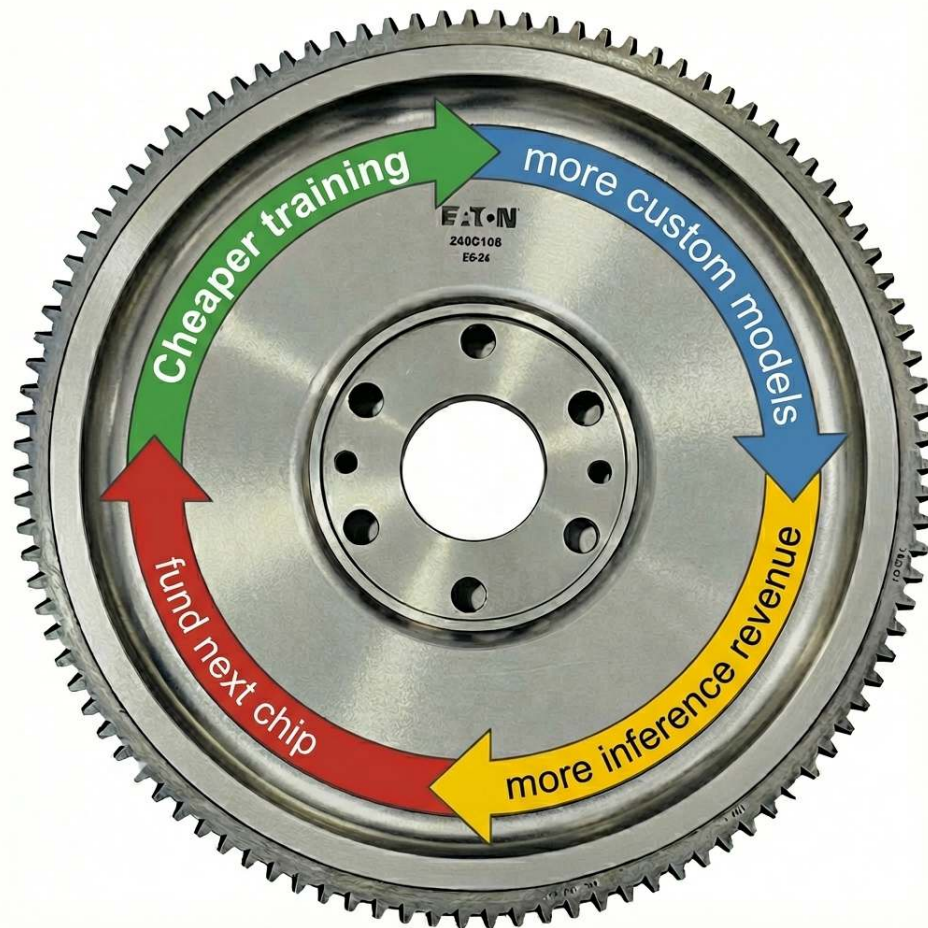
Nova Forge: The Price Tag

- \$100,000/year access fee
- Plus compute costs

Nova Forge: Who It's For

- Enterprises with big proprietary datasets
- Code, tickets, logs, documents, call transcripts
- Companies that can afford six-figure entry costs

The Margin Weapon: Trainium



Trainium: The Economics

- Trainium3 + Trn3 UltraServers
- 4x performance over previous gen
- Major energy and cost improvements

Trainium: Why It Matters

- Makes Nova Forge economically viable
- Compresses marginal cost of each training cycle
- Enables iterative specialization

Trainium: Provisioning

```
import pulumi_aws as aws

trn1_instance = aws.ec2.Instance("trn1-instance",
    instance_type="trn1.2xlarge",
    ami=ami.id,
    subnet_id=subnet.id,
    vpc_security_group_ids=[security_group.id],
    tags={"Name": "trn1-training-instance"},
)
```

The Data Moat Play

- For most companies, this whole stack is overkill.
- If your AI roadmap is "add a chatbot," you don't need Nova Forge.
- Hosted models + RAG + agents will get you 90% there.

When Baked-In Training Wins

- LLMs behave like the distributions they're trained on
- Your proprietary data in the core training loop means the model *behaves* like someone who's lived inside your systems
- That's different from "we stuffed a PDF into the context window"

The Performance Angle

- High-volume workflows: support triage, routing, code review, fraud checks
- "Generic model + giant prompt + RAG + tools" is slow and expensive
- A smaller, specialized model runs with smaller contexts, simpler prompts, fewer tool calls

Where the Models Work: AgentCore



AgentCore: What It Is

- Managed runtime for AI agents
- Sticky per-session microVM
- Built-in memory (short and long-term)
- Identity, permissions, observability

AgentCore: The Agent Code

```
from strands import Agent
from strands.models import BedrockModel

fraud_agent = Agent(
    model=BedrockModel(
        model_id="us.anthropic.claude-opus-4-5-20251101-v1:0"
    ),
    tools=[get_user_profile, get_recent_transactions,
           block_credit_card],
    system_prompt="You are a Senior Fraud Analyst..."
)
```

AgentCore: New at re:Invent

- **Policy** — guardrails
- **Evaluations** — quality checks
- **Episodic Memory** — per-session state

AgentCore: Deploying with Pulumi

```
import pulumi_aws as aws

agent_runtime = aws.bedrock.AgentcoreAgentRuntime(
    "fraud-detection-agent",
    agent_runtime_name="fraud_detection_agent",
    role_arn=role.arn,
    agent_runtime_artifact={
        "container_configuration": {
            "container_uri": f"{ecr_repo.repository_url}:latest",
        },
    },
    network_configuration={"network_mode": "PUBLIC"},
)
```

The Proof of Concept: Nova Act

- Browser-based UI automation
- Form filling, search-and-extract, QA testing
- Amazon claims ~90% reliability
- Deploys directly to AgentCore Runtime

Nova Act: Not Just "LLM + Playwright"

- Specialized Nova 2 Lite variant
- Trained on synthetic "web gym" environments
- RL loop: did the workflow succeed?
- Wrapped in AgentCore with tools and policies

One Stack to Rule Them All

- **Trainium** lowers training costs
- **Nova Forge** plugs in your data and rewards
- **AgentCore** is where models act
- **Nova Act** shows the pattern in action

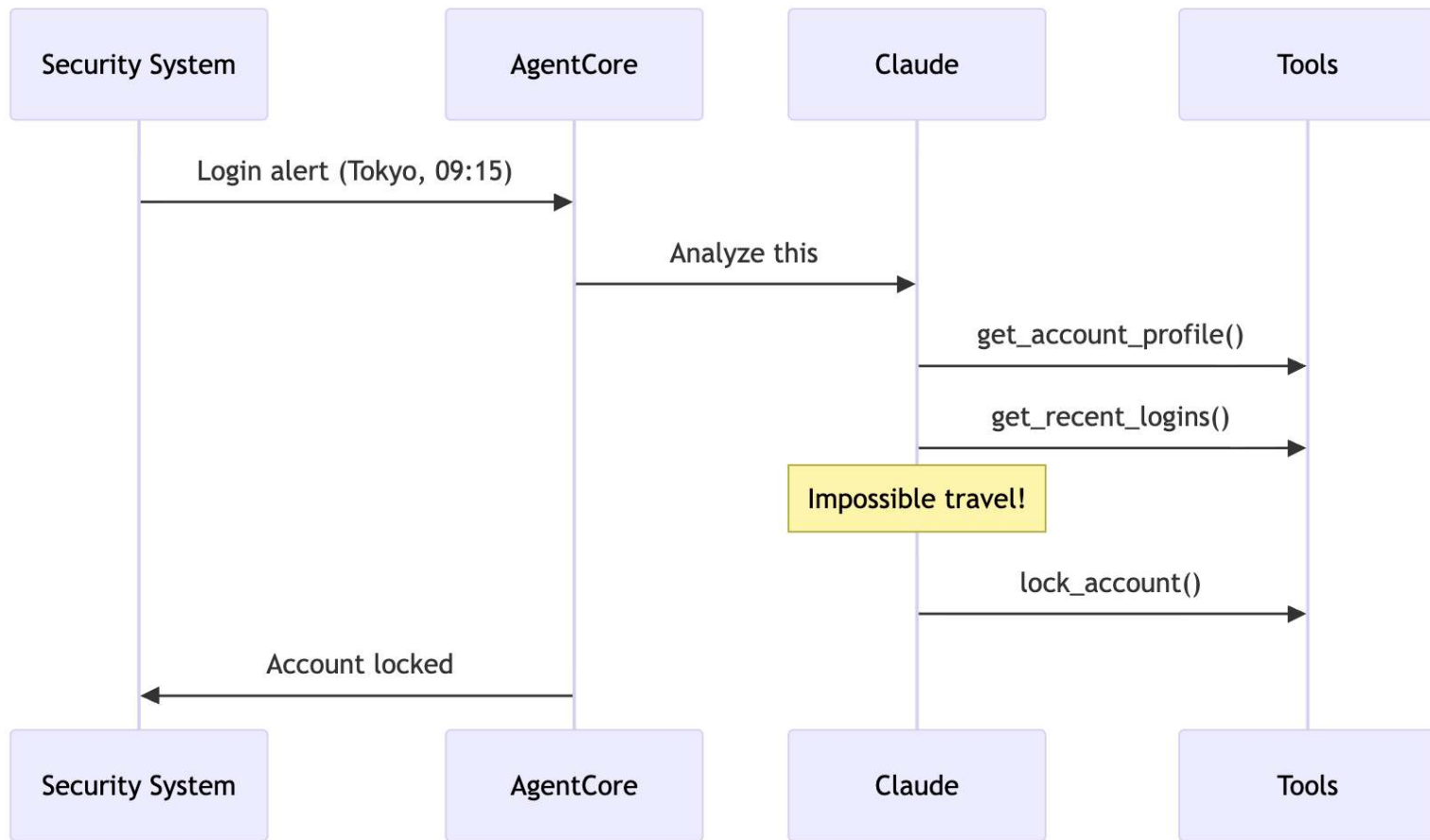
Who This Is For

- Enterprises with proprietary data worth baking in
- Reward loops that can guide specialization
- Operational maturity to run custom models

AWS's Bet

- Enterprise AI is moving past stock foundation models
- The future: **agents shaped by proprietary data and domain feedback**
- Most companies won't build the infrastructure
- So AWS is offering to rent them the whole pipeline

AgentCore: Login Security Example



Deploying AgentCore

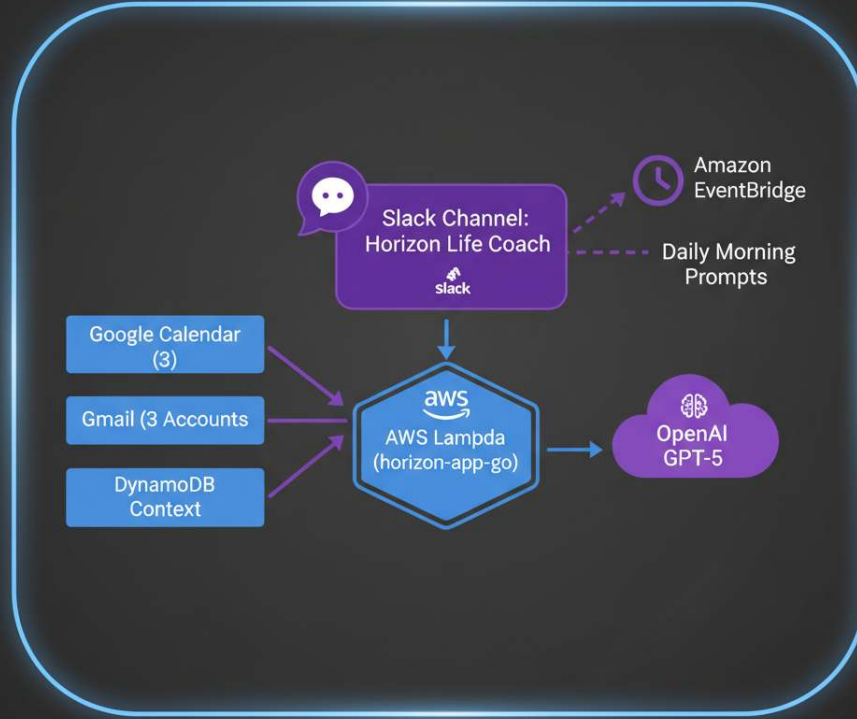
- How do you actually ship this?
- IAM roles, ECR, guardrails, runtime config
- Infrastructure as code

What Does Customized AI Look Like?

→ Slack Demo

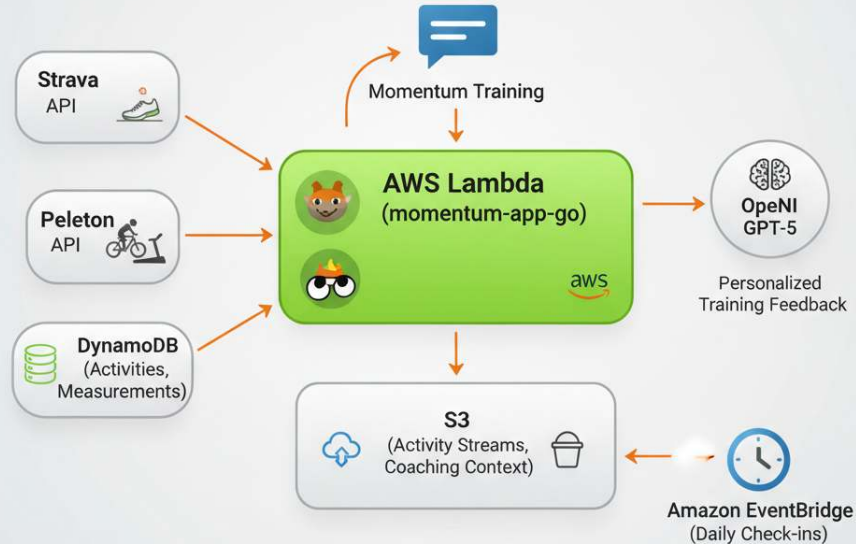
Horizon: Life Coach

Horizon: Your Personal AI Life Coach



Momentum: Training Coach

Momentum Training Coach Bot Architecture



Thank You!

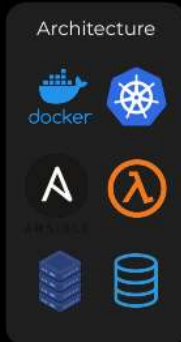
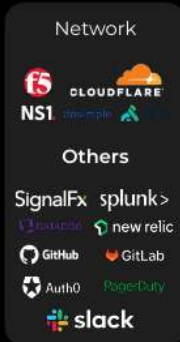
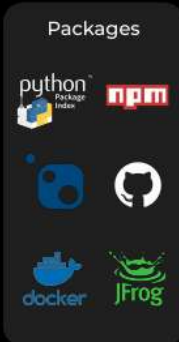
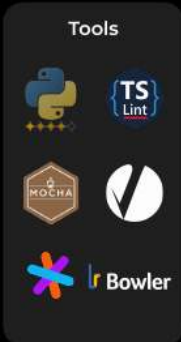
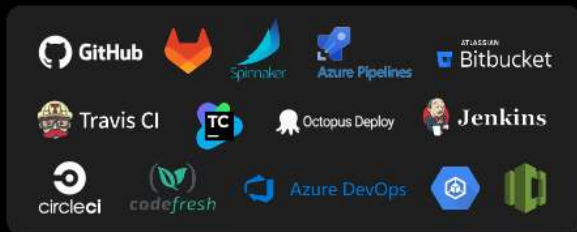
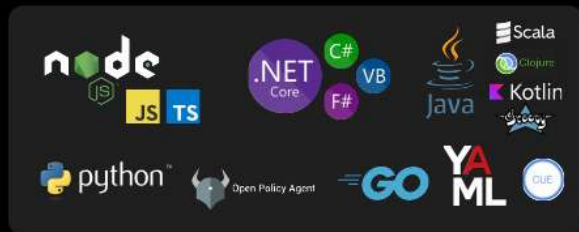
Adam Gordon Bell · Community Engineer ·
Pulumi.com



Appendix: Pulumi Overview

Pulumi IaC

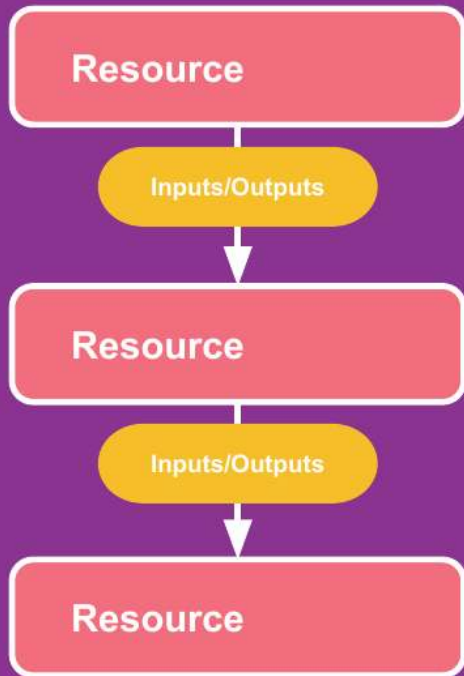
Your cloud, your language, your way



170+ Clouds Supported

Project

Program



Stack (DEV)

Stack (QA)

Stack (PROD)

Pulumi Program Model



Pulumi IaC Model

