Accionlabs

Generative AI in a Box

For cost-efficient, high-throughput Generative AI deployments

Top 3 Challenges in Gen AI Adoption

Addressing critical challenges in GenAI adoption, including concerns over data security, performance optimization, & the high costs of infrastructure & LLM usage

Data Security

Concerns over sensitive data exposure

Performance

Time-consuming optimization process

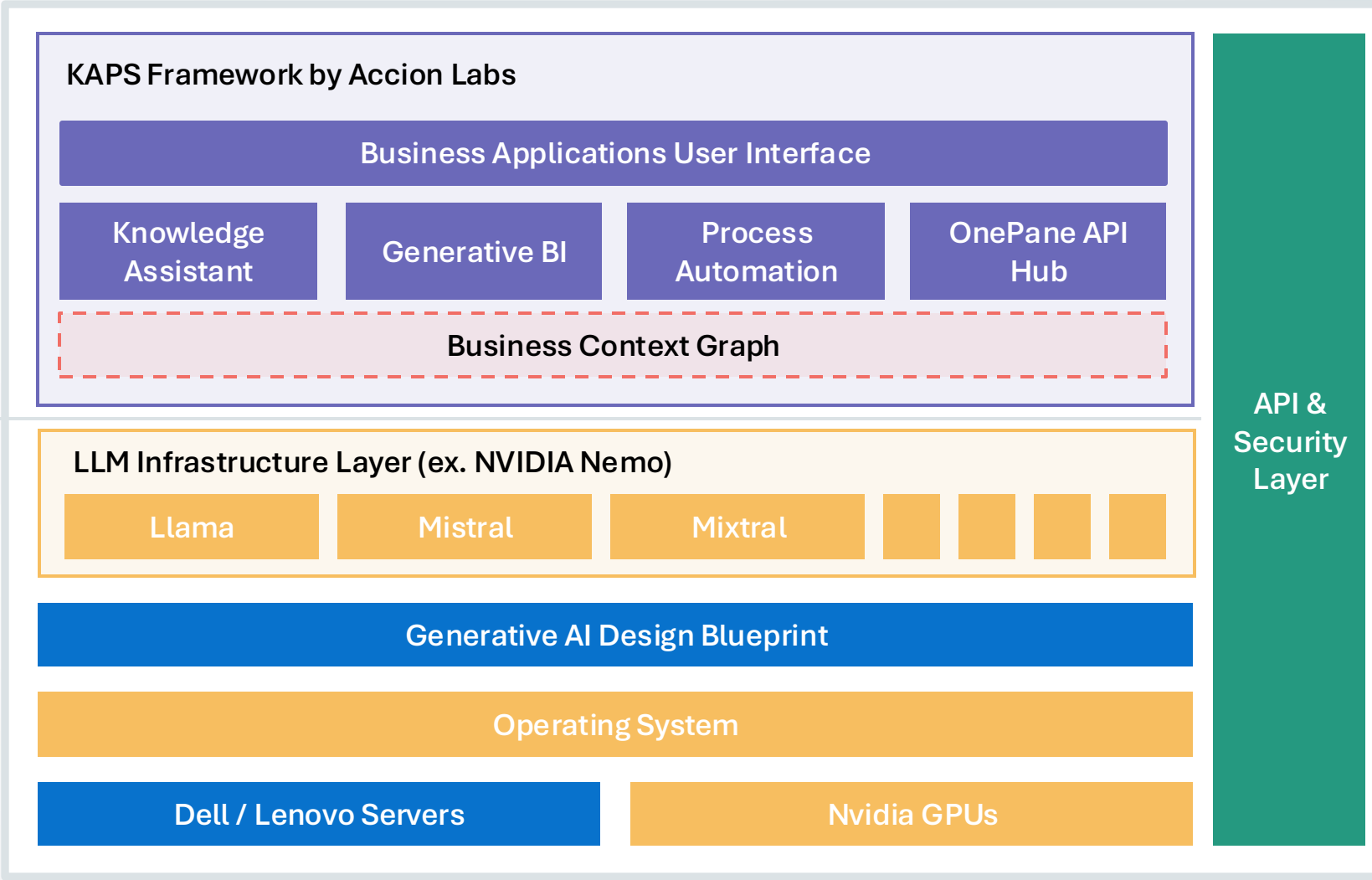
High Adoption Cost

Infrastructure and LLM usage costs

Gen AI in a Box Offering

Deployable apps and solution development framework - KAPS

AI platforms, pluggable models, design blueprints, Infrastructure managers and hardware



Overcoming Adoption Barriers with Gen AI in a Box

Cost Efficiency

- Optimized Hardware: Maximize AI performance with tailored hardware solutions
- On-Premise Hosting Option: Keep data local to reduce cloud costs
- Open-Source Models: Access advanced AI capabilities without licensing fees

Accuracy

- Task-Specific Models: Tailor Transformer or SSM models for specific operational needs
- Model Fine-Tuning: Customize models meticulously for enhanced precision
- Graph RAG Architecture: Enhance coherence using retrieval-augmented generation techniques

Security

- On-Premise Data: Maintain data privacy and control with local storage
- Access Control: Safeguard sensitive information with robust management
- Guardrails Configurations: Ensure safety and compliance with predefined AI behavior rules

Accelerators

- Gen AI Development Framework: Rapidly create generative AI solutions with a robust framework
- Ready-to-Deploy Applications: Quickly deploy pre-configured applications for immediate use
- Industry-Specific Solutions: Customized AI solutions to meet specific industry challenges

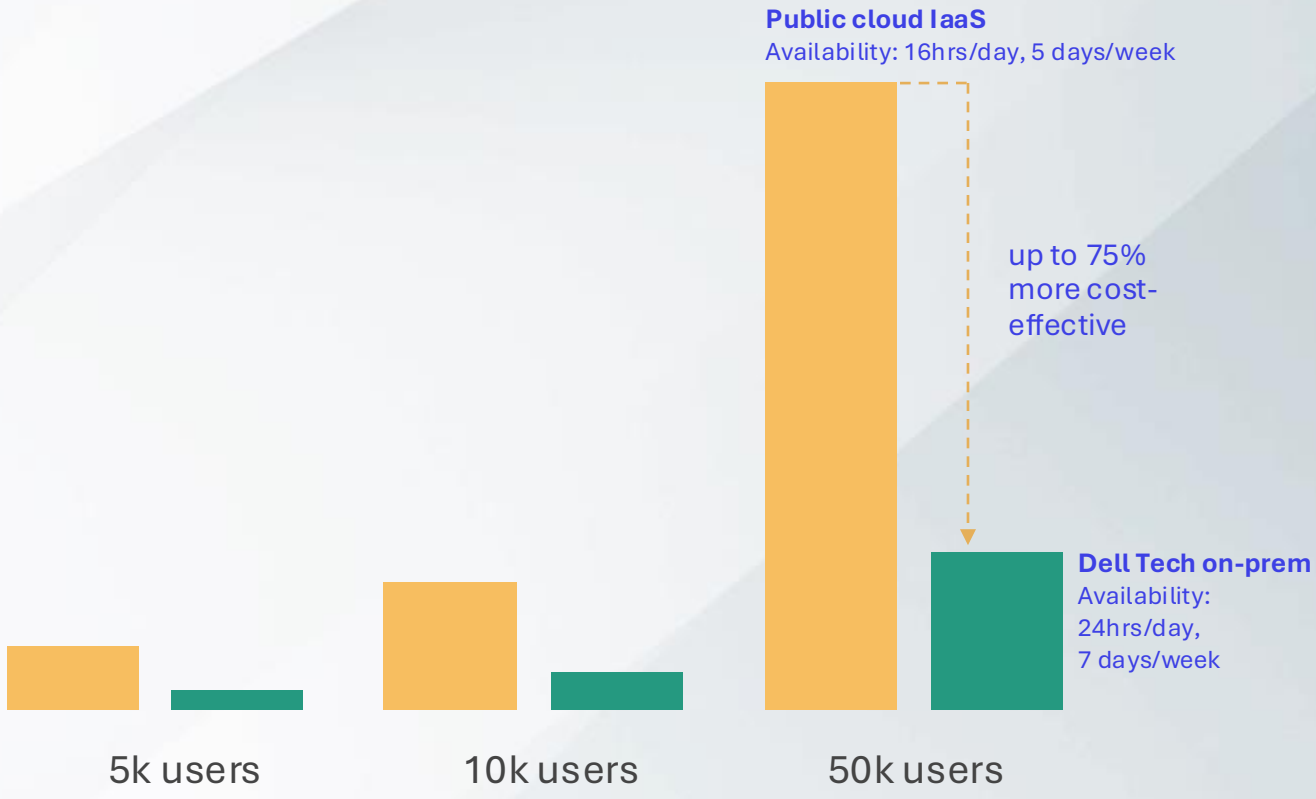
Cost is a major consideration for LLM inferencing

- Dell commissioned a study with Enterprise Strategy Group (ESG) comparing the expected costs to inference LLMs on-prem with Dell infrastructure vs. **public cloud IaaS** and **API services1.**
- Over a three-year period, Dell can provide inferencing that is up to:

75%
more cost-effective
than public cloud IaaS

88%
more cost-effective
than token-based API services

Expected cost to deliver Inferencing for
70B parameter Llama 2 LLM using RAG



1. Based on Enterprise Strategy Group research commissioned by Dell, comparing on-premises Dell infrastructure versus native public cloud infrastructure as a service, April, 2024. Analyzed models show a 7B parameter LLM leveraging RAG for an organization of 5k users being up to 38%more cost effective while a 70B parameter LLM leveraging RAG for an organization of 50k users being up to 75% more cost effective. Actual results may vary.Economic Summary

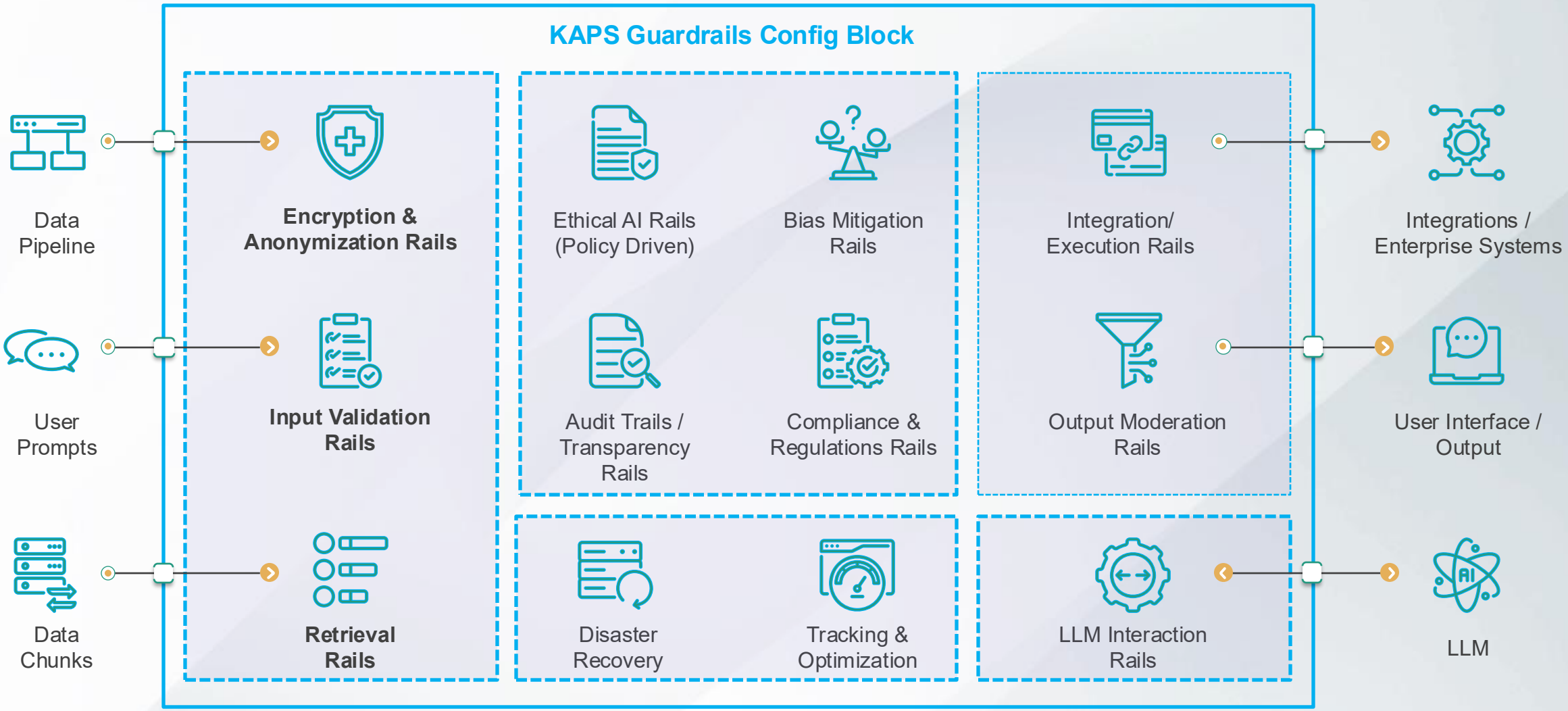
SLMs in the Box

Small Language Models (SLMs) are optimized AI models with fewer parameters compared to massive, general-purpose LLMs. They are designed for:

- **Efficiency** – Lower compute and memory footprint
- **Faster Inference** – Reduced latency for real-time applications
- **Cost-Effective** – Lower operational and hosting costs
- **Fine-tuned for Specific Use Cases** – Adapted for domain-specific knowledge

Why Use SLMs?	What We Use in "Gen AI in a Box"
Optimized Performance: Works well on-premise or edge devices	• **Llama** • **Mistral** • **DeepSeek** • **Qwen** • **NVIDIA NeMo** • **Other Domain-Specific Models**
Security & Compliance: Can be deployed in controlled environments	
Customization & Control: More adaptable for business-specific tasks	
Scalability: Ideal for enterprises that need AI without massive infrastructure	

Strategic Guardrails for Safe AI Operations



Respective cloud provider's solutions or open-source options like NVIDIA NeMo Guardrails can be used for implementations

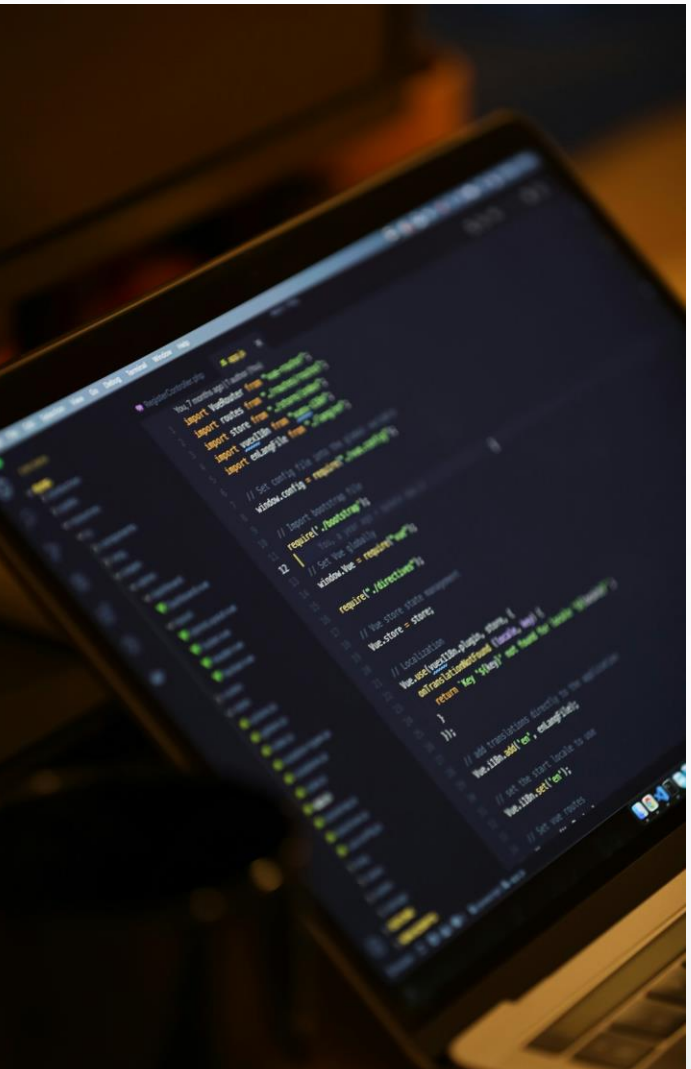
Accelerate GenAI outcomes and ensure long-term success with help at every stage

Establish strategy	Prepare data	GenAI Platform	Deploy & test model	Operate & Scale
Outcome Consensus on Roadmap	**Outcome** Validated data for model	**Outcome** Deployed GenAI platform	**Outcome** Tuned Model	**Outcome** Simplified GenAI operations
Accelerator Workshop Advisory Services	Services for Data Preparation, Data Security and Data Lakehouse	Deployment of GenAI in a Box Infrastructure by Partner Technologies Deployment of GenAI in a Box Software Stack by Accion	Use case implementation, test and improve accuracy and relevance Accelerator Services for RAG + on Precision workstations	Residency / or Managed Services from Partnered Technologies for GenAI in a Box Infrastructure

Simplify your AI journey with Accion and Partner Technologies

Services

Business Benefits



Accelerated Time To Market

- Ready-to-deploy business apps
- Base framework for quick custom solutions
- Shorter analysis & design cycles

Strategic Gen AI Adoption

- Strategic roadmap for Gen AI adoption
- Ready to explore use case portfolios
- Continuous capability enhancement

Data Security and Compliance

- On-premise deployment for control over data
- Compliance with data protection regulations
- Vital for sectors like healthcare and finance

Customization and Control

- Tailoring LLMs to specific needs
- Integration with existing systems
- Fine-tuning models for specialized tasks

Low Latency and Cost Control

- Achieving low latency for real-time applications
- High performance & accuracy, smaller footprint
- Cost savings in the long term

Finetuned LLM on Server

- Pre-fine-tuned domain-specific models
- Reduces manual annotation costs
- Creates highly performant, versatile LLMs

Accionlabs

Thank You

Generative AI in a Box concept not only resolves the most critical adoption challenges but also aims to revolutionize the integration of generative AI across every aspect of the ecosystem, encompassing infrastructure, technological frameworks, deployments, and ready-to-use business applications.