Domain AI (DSLM): The Next Enterprise AI Shift

Microsoft + Tech Mahindra —
Accelerating Enterprise AI Through Domain-Specific Language Models

Why Enterprise AI Is Moving Toward Domain Models

1

Cloud (2010s)

Infrastructure at scale

2

Data (2010s)

Analytics & warehousing

3

Generative AI(2022+)

Foundation models & APIs

4

Domain AI(2026 +)

Specialized intelligence at the operational core

Generic AI creates content → **Domain AI transforms operations.**

Generic AI Delivers

- Broad, general knowledge
- Open-ended responses
- Massive, expensive models
- Generic reasoning patterns

Enterprise Needs contextual Intelligence

Why most GenAI initiatives struggle to scale beyond the pilot stage.

Enterprise Requirements Demand

- Deep domain expertise
- Controlled, auditable outcomes
- Cost-efficient, right-sized models
- Regulatory alignment by design

⚠ **The missing layer is domain intelligence — purpose-built AI that understands your industry's language, rules, and workflows.**

Why Domain SLMs Are Emerging

Enterprise AI increasingly requires specialized models rather than larger models. The evidence is clear across every critical enterprise dimension.

Dimension	Generic LLM	Fine-Tuned LLM	Domain SLM
Accuracy	Low in context	Moderate	**High**
Inference Cost	Very High	High	**Optimized**
Compliance	Not designed for it	Partial	**Built-in**
Deployment Flexibility	Cloud only	Limited	**Edge / On-prem / Cloud**
Explainability	Black box	Partial	**Traceable**
Data Sovereignty	Shared infrastructure	Varies	**Sovereign by design**

Domain AI Creates a New Azure Consumption Engine

Every successful Domain SLM deployment creates a durable, long-term Azure workload — driving consumption across the full Microsoft stack.

Domain SLM

Purpose-built model trained on enterprise data

Azure AI Foundry

Model hosting, orchestration, and lifecycle

Azure OpenAI + Fabric

Inference, data pipelines, analytics integration

Security + Managed Services

Compliance, monitoring, and continuous operations

Every DSLM will become a long-term Azure workload.

Domain AI is not a one-time engagement — it is a recurring, expanding consumption motion that deepens the Microsoft relationship at every enterprise account.

What Makes Tech Mahindra Different

Our Differentiators

Build Models From the Ground Up

End-to-end model development from corpus curation to production deployment

Sovereign AI

Project Indus · Sahabat AI — proven sovereign model programs at national scale

Industry Expertise

Deep vertical capability in Telecom, BFSI, Energy, HLS and Manufacturing

Responsible AI

Governance, explainability, and compliance built into every model by design

The TechM DSLM Factory

powered through AXON – Joint Innovation Center

The Playbook - Capabilities

Discover

Identify high-value use cases with measurable ROI and domain data availability

Evaluate

Benchmark candidate models across accuracy, cost, latency, and compliance

Train

Fine-tune on curated domain corpus with SME knowledge integration

Validate

Rigorous evaluation against domain benchmarks and regulatory requirements

Deploy

Production deployment on Azure AI Foundry with enterprise-grade security

Operate

Continuous monitoring, retraining, and performance optimization at scale

TechM – Internal Teams aligned

Telecom • Network Operations • Customer Care	**Banking** • Payment Operations
Healthcare • Patient Care	**Lifesciences** • Patient Safety

**The above areas of exploring & building the DSLM has been initiated – with more to be identified and finalized with the joint teams involved in the DSLM plan*

DSLM Lifecycle: Building Domain Intelligence on Microsoft

To build, evaluate, and deploy domain-specific language models on the Microsoft stack — from raw data fabric to production-ready Copilots and Agents.

1

Business Domain

DSLM Blueprint by Industry — Define use cases, compliance boundaries, and domain-specific knowledge requirements that shape the entire model

2

Enterprise Data Fabric + OneLake

Unify structured and unstructured enterprise data across silos into Microsoft Fabric's OneLake — the governed, scalable foundation for model training and

3

Model Selection

Choose from the open and proprietary model ecosystem: **GPT**, **Phi**, **Llama**, **Mistral**, or **DeepSeek** — matched to domain complexity, latency, and cost

4

Azure AI Foundry

Evaluate, fine-tune, and benchmark candidate models. Apply safety guardrails and responsible AI policies before promotion — all within a governed MLOps

5

Domain SLM

A fine-tuned, domain-grounded Small Language Model optimized for precision, cost efficiency, and regulatory compliance within specific industry vertical.

6

Copilots · Agents · Business Apps

Deploy intelligence at the point of work — as embedded **Copilot** experiences, autonomous **Agents**, or integrated **Business Applications** that deliver measurable enterprise value.

This lifecycle is adapted for **Microsoft Azure stack** — ensuring enterprise-grade security, scalability, and seamless integration with existing investments in M365, Dynamics, and Azure services.

Evidence from Customer Deployments

Domain AI is delivering measurable business outcomes across industries. These are production cases — not proofs of concept.

30%

Faster Resolution

Telecom — Accelerated trouble-ticket analysis and network fault diagnosis

50%

Faster KYC

Banking — Reduced Know Your Customer processing time with full audit trail

<5s

Knowledge Retrieval

Energy — Sub-5-second engineering knowledge retrieval across decades of documentation

Case Study: Payments Operations DSLM

A leading financial institution replaced a costly GPT-based Text-to-SQL solution with a purpose-built Payments Domain SLM — achieving full compliance readiness in five months.

Before

- GPT-4 Text-to-SQL — high inference cost
- Non-deterministic query routing
- Significant audit risk and traceability gaps
- No compliance framework alignment

After

- Payments DSLM — optimized inference cost
- Deterministic, rule-governed routing
- Full end-to-end traceability by design
- Compliance-ready from day one

5-month transformation from pilot to enterprise-grade production AI.

Case Study: Telecom Network Operations DSLM

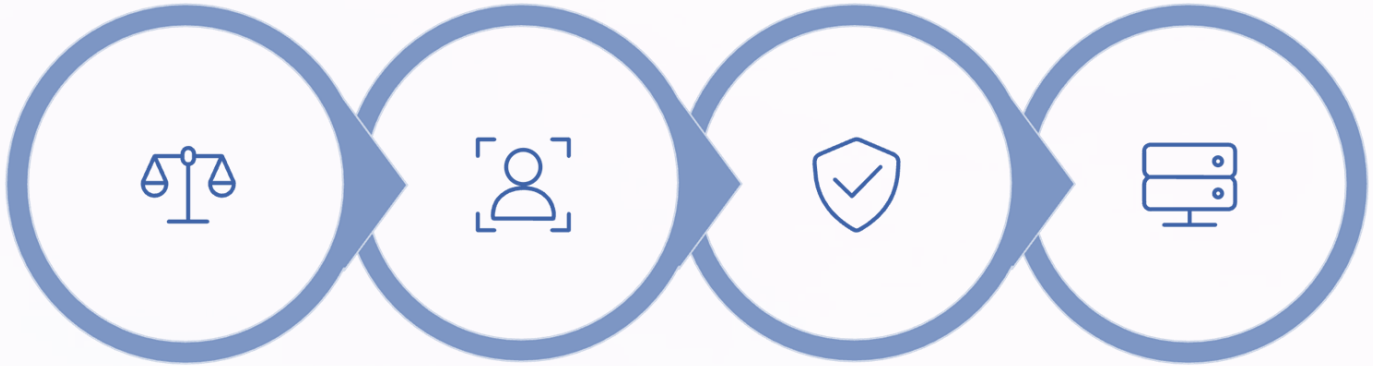
Built for CPU deployment at the network edge — no GPU infrastructure required. A compact, high-performance model that outperforms general-purpose alternatives on telecom-specific tasks.

85%	1.7B	40%
Resolution Accuracy	Parameter Model	Performance Gain
Automated network fault resolution with near-human expert accuracy	Compact by design — deployable on edge infrastructure without GPU dependency	Improvement over generic LLM baseline on network operations benchmarks

A smaller model that knows more than a larger one — because it was built for this domain.

Case Study: Sovereign Legal Intelligence Engine

Model selection driven by evidence, trust, and compliance . This engagement demonstrates TechM's sovereign-first methodology.



Evaluate Models

Fine-Tune Corpus

Govern Compliance

Deploy Sovereign

Candidate Models Evaluated

- DeepSeek
- Qwen
- GPT-OSS

Sovereignty first. Model selection based on evidence, trust, and compliance — not vendor preference.

How We Select the Right Domain Model

Evaluation First | Model Second

Evaluation Dimensions

- Domain Accuracy
- Latency
- Inference Cost
- Data Sovereignty
- Explainability
- Compliance Fit

Candidate Model Pool

We evaluate across the full landscape — to meet customer outcome

- GPT
- Phi
- Claude
- Llama
- Qwen
- DeepSeek
- Mistral
- Gemini

Model selection is evidence-led . The right model is the one that performs best in customer domain

Scalable Industry AI Patterns

Pre-built, validated domain AI solutions that compress time-to-value, reduce delivery risk, and drive greater Azure consumption from day one.

Banking — Payments Intelligence

Deterministic payment routing, fraud detection, and KYC acceleration with full audit traceability

Telecom — Network Operations

Edge-deployable DSLM for fault resolution, NOC automation, and predictive network management

Energy — Engineering Knowledge Assistant

Instant retrieval of engineering documentation, standards, and operational procedures

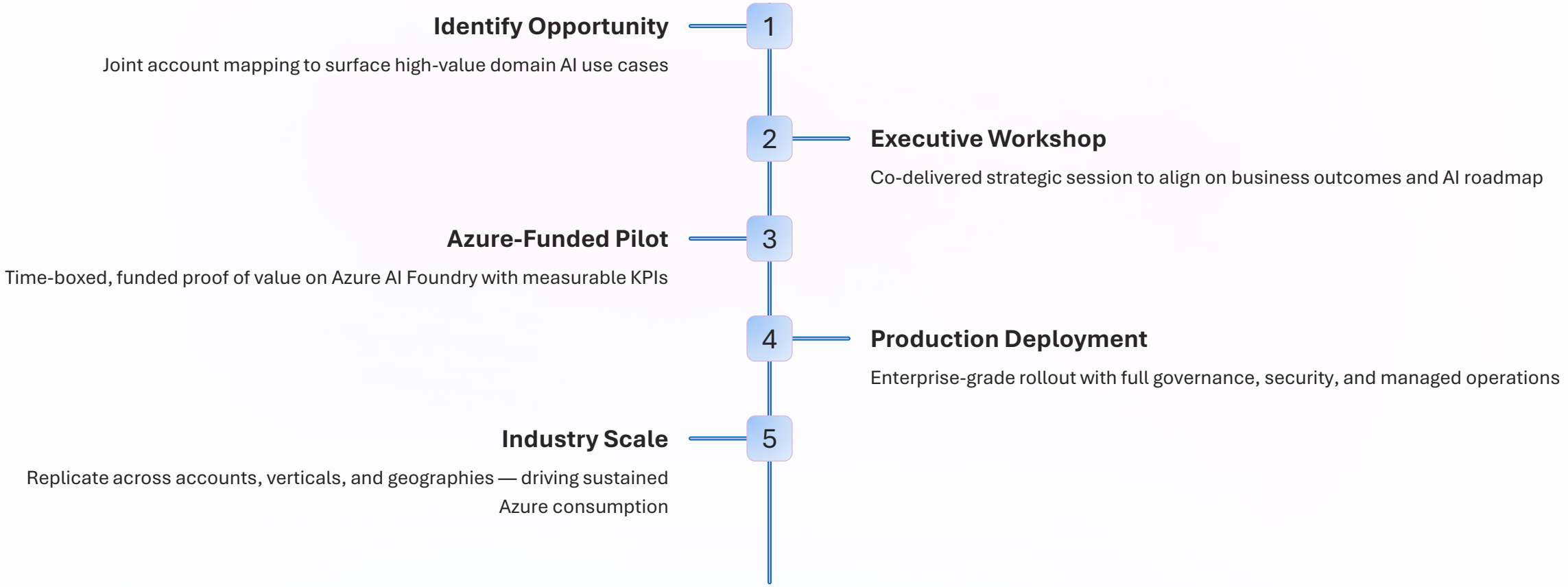
Public Sector — Sovereign AI Assistant

Compliant, sovereign AI for citizen services, legal intelligence, and policy analysis

Open : Operating Model along with Microsoft Team to identify the focus area – vertical – geo mix & work to drive engineering and GTM with AXON – DSLM Track

Joint Microsoft + TechM GTM Motion

A structured, repeatable co-sell motion designed to accelerate enterprise AI adoption — from first conversation to industry-scale deployment.



Repeatable model — **Identify → Workshop → Pilot → Deploy → Scale.**

A Joint Opportunity to Define & Build Domain AI – Focus Verticals & Use Cases

Create a durable, differentiated market position for both organizations.

Microsoft Brings

- Azure platform, scale, and global reach
- AI Foundry, OpenAI integration, and Fabric
- Enterprise customer relationships and trust

Tech Mahindra Brings

- Domain expertise and model engineering depth
- Industrialized DSLM delivery capability
- Proven production deployments across verticals

Next Steps (Discussion & Approval)

DSLM Factory – MS Team	Target Accounts
Azure Funding	Joint Workshops

Thank You

Appendix

Reference materials supporting the Domain AI thesis — sovereign programs, architecture, frameworks, and validated industry deployments.

Project Indus

India's first multilingual large language model — built by TechM for sovereign AI at national scale, supporting 22 Indic languages

Sahabat AI

Southeast Asia's sovereign AI initiative — a family of models purpose-built for Bahasa Indonesia and regional language contexts

DSLM Reference Architecture

End-to-end Azure-hosted architecture covering data ingestion, model training, inference, and enterprise integration patterns

Responsible AI Framework

TechM's governance-by-design methodology — covering fairness, transparency, accountability, privacy, and regulatory compliance

Training Pipeline & Industry Use Cases

Detailed training pipeline documentation and validated use case library spanning Banking, Telecom, Energy, and Public Sector verticals

Project Indus – The first Hindi LLM built grounds up by Tech Mahindra

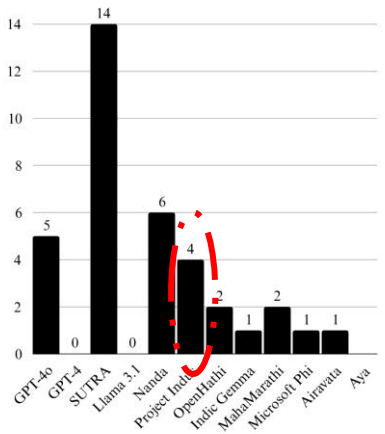


Figure 3: Number of Best Performances Achieved by Each Tokenizer Across 22 Languages.

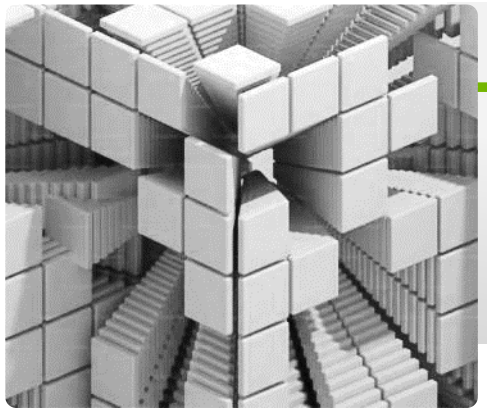
Small Models can work well

GPT-4	SUTRA	Llama 3.1	Nanda	Project Indus	OpenHathi	Indic Gemma	MahaMarathi	Microsoft
1.4	**0.4571**	1.4	1.4	2.7714	1.5714	0.8286	1.3143	1.5428
1.2307	**0.2115**	1.25	1.25	2.8076	1.3461	0.5769	1.0961	1.3269
1.0540	0.5405	0.5945	**0.4594**	0.4864	0.5405	0.5675	0.5405	1.2432
1.0313	0.4688	0.5938	**0.3750**	0.4688	0.4063	0.4688	0.4063	1.0312
1.6875	**0.4688**	1.7188	1.7188	2.75	2.6875	0.7188	1.0938	2.6562
1.0	0.4545	0.5909	0.3636	**0.3182**	0.4545	0.5455	0.3636	1.0454
1.76	**0.44**	1.8	1.8	2.84	2.52	0.56	1.12	2.48
1.093	**0.5814**	0.8837	0.8837	1.8605	1.1628	0.5814	1.186	1.1395
1.1429	0.5357	0.6429	0.5	0.4643	0.6071	0.5357	1.1071	0.6071
1.0	0.6316	0.6316	**0.3684**	**0.3684**	0.5789	0.5789	1.1578	0.5789
1.75	**0.5**	1.8333	1.8333	3.0	1.3333	0.6667	1.25	1.3333
1.2941	**0.5882**	1.3529	1.3529	2.8824	1.5882	0.7647	1.5882	1.5294
0.9412	0.5294	0.6471	**0.3529**	**0.3529**	0.4706	0.5882	1.0582	0.4706
0.9091	**0.3182**	0.6364	**0.3182**	0.3636	**0.3182**	0.4545	0.4090	1.1363
2.625	**0.625**	2.625	2.625	2.875	2.875	1.0625	2.875	2.8125
1.6923	**0.4615**	1.7308	1.7308	2.7692	2.3077	0.7692	2.3077	2.2692
1.0833	0.6667	0.5833	**0.5**	**0.5**	**0.5**	0.6667	1.08333	**0.5**
2.647	**0.4705**	2.7058	2.7058	2.8823	2.8823	1.0588	2.9411	2.8235
0.9117	0.5	0.6176	0.6176	1.8529	1.0882	0.5588	1.1176	1.0588
1.3823	**0.3823**	1.4117	1.4117	2.7647	1.2352	0.5294	1.0882	1.2058
1.75	**0.2916**	1.7916	1.7916	2.8333	2.6666	0.625	1.125	2.625
0.7857	**0.3571**	0.5357	1.8928	1.8928	1.1071	0.4285	1.1071	1.0714

Table 2: Average NSL Values Across Models for 22 Languages (lower is better). The bold values indicate the best-performing tokenizer for each language.

- Fine Tuned Indus model of 1.2B parameter model deployed on Intel 5th Gen Powered platform for Inference serving
- In-Box Scale out using the Sub NUMA cluster has been applied
- Inference throughput found to be ~35 tokens /s and ttft for 100 concurrent requests is ~10s
- Published Whitepaper jointly: https://cdrdv2-public.intel.com/830573/white-paper-benchmarking-indus-language.pdf

Indonesia LLM: The road ahead



Sahabat-AI

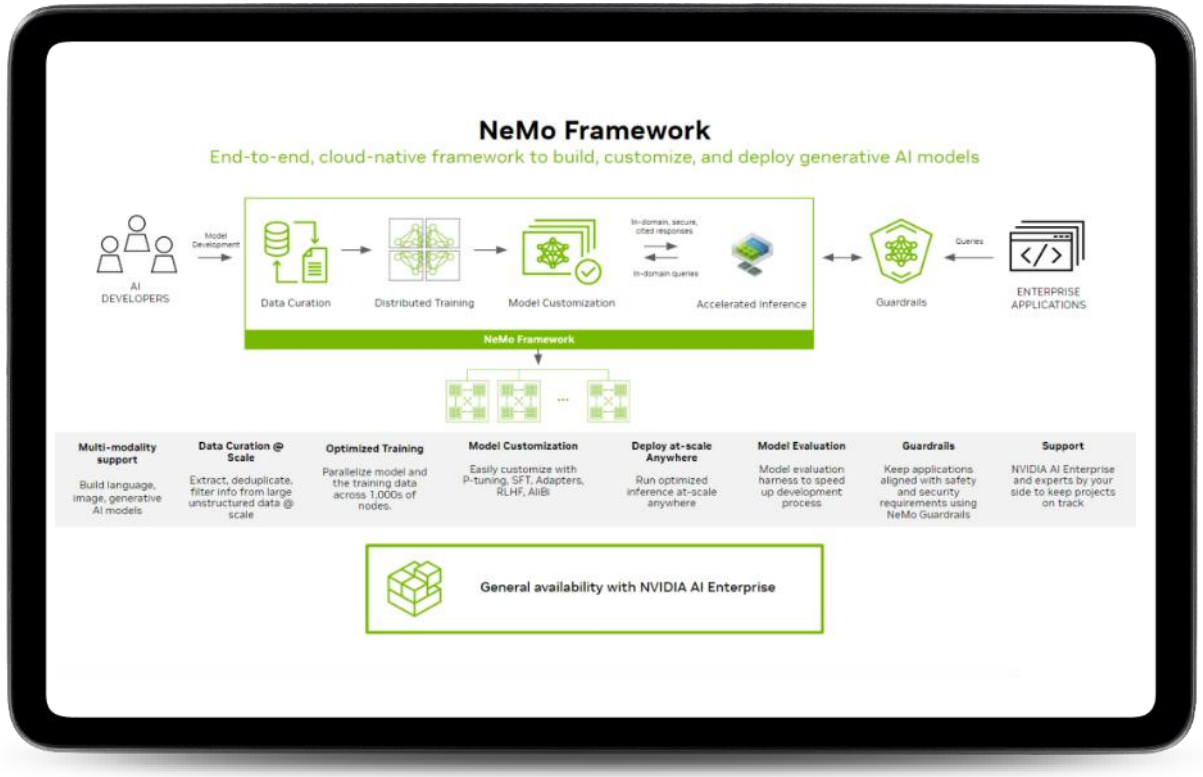
~8 bn
Parameters

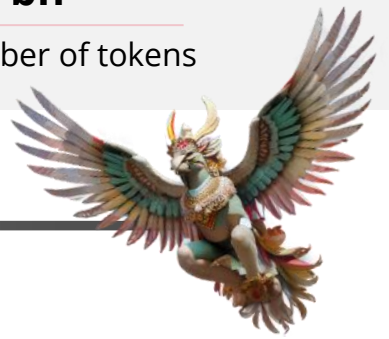
H100
GPUs

6 GB
Model fine-tuning data

100+ GB
Model training data

150 bn
Number of tokens

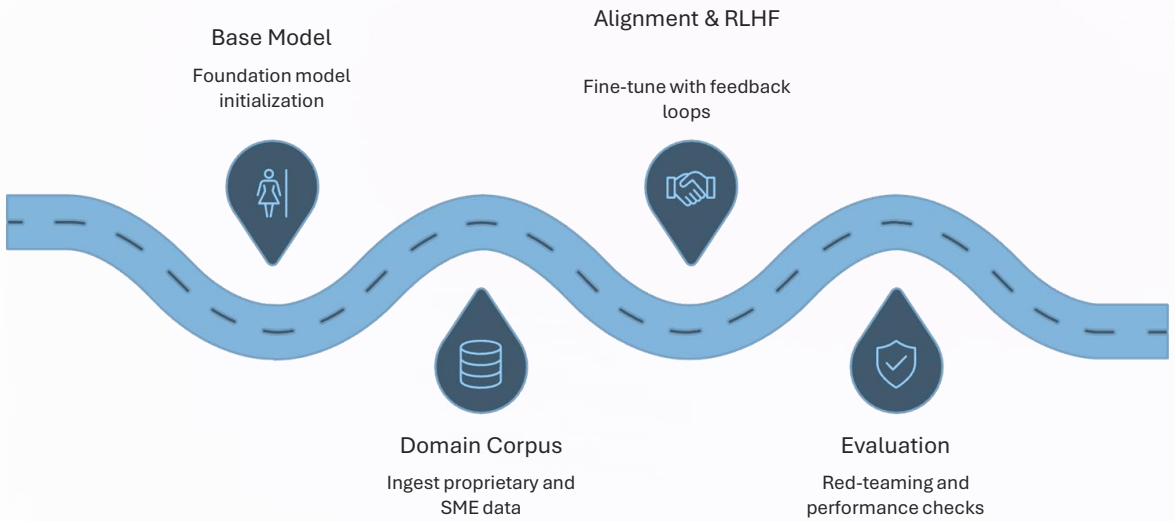




Model	Language	Score from Web (Huggingface)	Benchmarks	
			Hendrycks Repo	EleutherAI Repo
Llama 3 8B	English	0.684	0.739	0.681
	Bahasa	N/A	0.651	0.603
GPT 3.5	English	0.689	0.711	0.683
	Bahasa	N/A	0.685	0.611
SeaLLM	English	0.618	0.649	0.619
	Bahasa	N/A	0.815	0.689
Sahabat AI	**English**	**N/A**	**0.712**	**0.686**
	Bahasa	**N/A**	**0.803**	**0.704**

We are state of the art

From Foundation Models to Domain Experts



What Goes Into the Domain Corpus

Domain Data

Proprietary datasets curated for the target industry vertical

SME Knowledge

Subject matter expert annotation, validation, and feedback loops

Operational Logs

Real-world transaction and process data from enterprise systems

Policies & Regulations

Regulatory frameworks, compliance requirements, and internal procedures