FOR INSURANCE LEADERS · SUBMISSION, UNDERWRITING, CLAIMS, SERVICE & INSIGHTS

Build an Adaptive Domain Language Model Farm for Insurance

InsightDLM is an end-to-end framework from VerticalServe to curate your insurance data, fine-tune **Domain Language Models (DLMs)** — both **Small Language Models (SLMs)** and **Large Language Models (LLMs)** — and stand them up as an **adaptive model farm** alongside frontier LLMs (Claude, OpenAI) served through **Bedrock or Azure secure inference**. Purpose-built for the full insurance value chain: **submission & distribution, underwriting, policy servicing & operations, claims, customer 360 insights, and compliance reporting**.

The architecture: a smart router that picks the right model per request — small domain SLMs handle high-volume bounded work at the lowest cost and tightest SLAs, domain-tuned LLMs handle complex insurance reasoning and long-form drafting, and Bedrock / Azure-served frontier LLMs (Claude, OpenAI) handle the broad, novel reasoning that SLMs are *not* built for. All inside your security boundary, all auditable, all yours to operate.

10–100×
Lower cost per call when SLM-tier DLMs absorb high-volume bounded work

<600 ms
Target p95 latency for in-workflow agent and underwriter assist

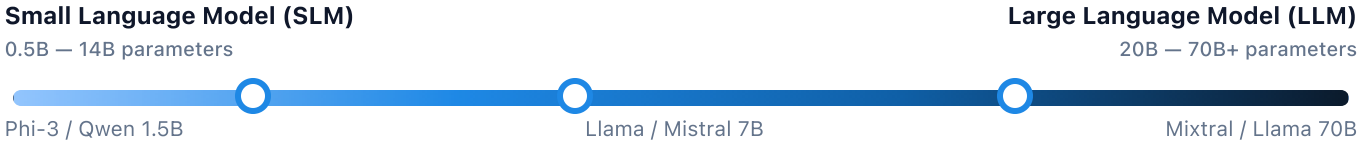
0
Data leaving your security boundary — VPC, on-prem, or Bedrock/Azure

THE CONCEPT

A DLM is a language model purpose-built for *your* domain

A **Domain Language Model (DLM)** is a language model trained or fine-tuned on the corpora, terminology, forms and workflows of a specific industry — here, insurance — and deployed inside the customer's own security boundary. A DLM is defined by **what it knows, how it is evaluated, and where it runs**, not by its parameter count.

A DLM can be a **Small Language Model (SLM)** or a **Large Language Model (LLM)**. InsightDLM lets you build a portfolio of both sizes and operate them together with frontier models — an **adaptive model farm** (described on the next page).



When an SLM-tier DLM is the right tool

- **High-volume, well-bounded tasks** — submission triage, ACORD extraction, FNOL classification, intent routing
- **Tight latency budgets** — in-workflow agent & underwriter assist
- **Edge or on-device** — field-adjuster laptops, broker portals
- **Cost-sensitive volumes** — millions of calls per month

When an LLM-tier DLM is the right tool

- **Complex multi-document domain reasoning** — coverage analysis across endorsements, multi-policy 360 narratives
- **Long-form generation on your data** — UW referral memos, regulatory filings, denial letters that hold up to legal review
- **Lower-volume, higher-stakes work** on insurance-specific content
- **Where domain quality outweighs cost** per call

DLM VS. OTHER APPROACHES

How a DLM is different from generalist LLMs and RAG-only stacks

Generalist LLM (Claude, GPT)
Trained on the open web. No knowledge of your endorsements, submissions, claim notes or operational playbooks. Strong at broad reasoning, weak on insurance specificity.

RAG-only over a generalist LLM
Retrieves your documents at inference, but the underlying model still doesn't *speak* insurance. Cost and latency belong to the hosted LLM. Hallucinations on policy language remain a risk.

Domain Language Model (DLM)
Fine-tuned on your insurance corpora. Speaks the domain natively. Combined with retrieval where useful. Deployed in your VPC or on-prem. Sized SLM or LLM per task. Owned by you.

The conclusion most insurers reach: there is no single right model. The right answer is a **portfolio** — SLM-tier DLMs for bulk traffic, LLM-tier DLMs for complex domain work, and frontier LLMs (Claude, OpenAI) via Bedrock or Azure secure inference for broad reasoning the DLMs aren't built for. InsightDLM is the framework that builds, evaluates, deploys and operates that portfolio as one system. See page 3.

THE PRODUCTION ARCHITECTURE

An adaptive model farm routes each request to the right model

An **adaptive model farm** is the production architecture InsightDLM stands up for you: a portfolio of language models served behind an **intelligent routing layer** that picks the right model per request based on **task complexity, confidence from upstream tiers, latency SLA, cost budget, and data sensitivity**. SLM-tier DLMs absorb the bulk of bounded, high-volume insurance traffic at the lowest cost. LLM-tier DLMs handle complex reasoning and long-form generation on your data. Frontier LLMs (Claude, OpenAI) served via **Bedrock or Azure secure inference** handle broad, novel reasoning that DLMs are not built for.

ADAPTIVE ROUTING LAYER

Confidence · Task Complexity · SLA · Cost · Data Sensitivity · Volume Class

A single API surface for every insurance workflow — the right model is chosen per call

TIER 1 · SLM DLMS

Owned Small Domain Models

0.5B – 14B params · Qwen / Llama / Mistral / Phi / Gemma

BEST FOR

- Submission triage & ACORD/SOV extraction
- FNOL classification & severity
- Intent routing, complaint classification
- Short-context summarization, endorsement triage

OPERATING PROFILE

Latency: <500 ms p95

Cost: ~$0.02–$0.20 per 1K calls

Serving: vLLM / SGLang in your VPC or on-prem

Footprint: single A10/L4/L40S class GPU

TIER 2 · LLM DLMS

Owned Large Domain Models

20B – 70B params · quantized for cost-effective serving

BEST FOR

- UW referral memos & appetite analysis
- Multi-policy coverage analysis
- Customer / commercial-account 360 narratives
- Regulatory filing drafts on your data
- Long-context claim file synthesis

OPERATING PROFILE

Latency: ~1–3 s typical

Cost: ~$0.20–$1.00 per 1K calls

Serving: vLLM with GGUF/AWQ/GPTQ quantization

Footprint: 1–2 A100/H100 class GPUs

TIER 3 · FRONTIER LLMS

Claude / OpenAI via Bedrock / Azure

Secure inference inside your cloud account — no model training on your data

BEST FOR

- Broad, novel reasoning DLMs aren't trained for
- Exploratory analytics & executive research
- Edge-case escalations from Tier 1/2
- One-off creative drafting where breadth wins

OPERATING PROFILE

Latency: ~1–5 s typical

Cost: ~$5–$30 per 1K calls

Inference: Bedrock / Azure OpenAI in your account

Controls: customer-managed keys, DPA/BAA, audit logs

What custom SLMs are *not* for. SLM-tier DLMs are powerful where the task is bounded and well-defined. They are **not** a good fit for: open-ended reasoning beyond the training distribution, multi-step chain-of-thought across unfamiliar domains, queries that need the latest world knowledge, one-off exploratory analyst questions, or long-form creative drafting where breadth matters more than precision. For these, the router escalates to a Tier 2 LLM DLM (if it's on-domain) or to Tier 3 (Claude / OpenAI via Bedrock or Azure).

Why Bedrock / Azure for frontier LLMs. When frontier reasoning is needed, InsightDLM routes to Claude or OpenAI through **Bedrock or Azure OpenAI secure inference**, not direct hosted APIs. This keeps inference inside your existing cloud account boundary, runs under your IAM / VPC controls, supports customer-managed keys (KMS / Key Vault), is covered by enterprise DPAs/BAAs, and emits to your audit trail — with no model training on your data.

THE MISSION

Make it practical to *own* a fine-tuned DLM portfolio — not rent a generalist API

InsightDLM gives insurers an opinionated, end-to-end framework to curate proprietary data, fine-tune the right open-weight base model (Qwen, Llama, Mistral, Phi, Gemma; 0.5B – 70B+), evaluate it against insurance-specific scorecards, and serve it inside your environment alongside Bedrock and Azure-hosted frontier LLMs — with full lineage and audit trails.

THE LIFECYCLE

Four reproducible phases

1

Curate

Ingest policies, submissions, ACORD forms, claim notes, call transcripts, agent emails, CRM/PAS/claims tables. Parse, OCR, dedupe, scrub PII/PHI, classify, version with lineage.

2

Synthesize & Label

Generate domain Q&A, instructions, reasoning traces and hard negatives from your corpora using teacher-LLM distillation and human-in-the-loop labeling.

3

Train & Evaluate

Fine-tune with reusable recipes — SFT, LoRA / QLoRA, DPO / ORPO, continued pretraining. Score against domain eval suites and red-team probes; regression-gate releases.

4

Deploy & Manage

Quantize (GGUF / AWQ / GPTQ / MLX). Serve via vLLM / SGLang / TGI / llama.cpp. Guardrail SLMs gate output. Registry tracks drift, cost and quality across the model farm.

THE THREE INTEGRATED PLANES

What InsightDLM gives you out of the box

Data Curation Pipelines

Layout-aware document parsing, OCR fallback, PII/PHI scrubbing, deduplication, quality filters, synthetic Q&A generation, versioned datasets with full source-to-hash lineage.

Training Studio & Recipes

Base-model library (Qwen, Llama, Mistral, Phi, Gemma). SFT, LoRA / QLoRA, DPO / ORPO / KTO, continued pretraining, teacher distillation. YAML recipes versioned with data. Axolotl, TRL, Unsloth, DeepSpeed, FSDP.

Model Farm Ops & Routing

Domain eval harness with LLM-as-judge. Quantization (GGUF / AWQ / GPTQ / MLX). vLLM / SGLang / TGI / llama.cpp. Guardrail SLMs. Adaptive router with confidence/SLA/cost signals. Registry with drift & cost monitoring.

WHY GENERALIST LLMS ALONE FALL SHORT ON INSURANCE

Coverage & form hallucinations

Misread endorsements, exclusions, ACORD fields, state-form language.

Unpredictable cost at volume

Per-token meters become unbounded across submission, claims and service traffic.

Variable tail latency

Agent-assist and UW workbench need sub-second responses, not API timeouts.

PII / PHI exposure

Submissions, claim notes, medical records on a hosted API trigger reviews and DOI questions.

NOT JUST CLAIMS

Where an InsightDLM model farm pays off — end to end

Each box below is a cluster of DLM-fit tasks. Tier 1 (SLM DLMs) typically serves high-volume bounded work; Tier 2 (LLM DLMs) serves complex reasoning, long-form generation and 360 narratives; Tier 3 (Bedrock/Azure frontier LLMs) handles the rare broad-reasoning escalations. All routed by the adaptive router, governed by the same eval harness and registry.

DISTRIBUTION & SUBMISSION

New business intake at broker / agent volume

- Broker submission triage, summarization and prioritization
- Appetite matching & account clearance against UW guidelines
- ACORD 125 / 126 / 140 and SOV ingest at submission
- Producer / agent Q&A on appetite, forms, eligibility
- Auto-drafted quote, decline and follow-up letters

UNDERWRITING

Faster, more consistent risk decisions

- Risk appetite Q&A grounded in UW guidelines
- Exposure summarization from SOVs and engineering reports
- Loss-run analysis and loss-history narratives
- Referral memo drafting (account → senior UW)
- Renewal narratives and retention rationale

POLICY SERVICING & OPERATIONS

Cut handle-time on every service interaction

- Endorsement request triage & structured extraction
- Billing & payment inquiry classification + response drafts
- Coverage / renewal Q&A grounded in current policy set
- Mid-term adjustment intent and impact analysis
- SOP automation and ops knowledge Q&A for back-office

CLAIMS

Accuracy where it costs the most

- FNOL classification & severity scoring
- ACORD-25/27/125/140 extraction with field confidence
- Claim file, note & call summarization with next actions
- Coverage Q&A grounded in the policy in force
- SIU (fraud) & subrogation flagging with reasons
- Plain-language denial and status letter drafts

CUSTOMER 360 & INSIGHTS

A single, defensible narrative per customer

- Policyholder / commercial-account 360 summaries across policies, claims, calls, emails
- NPS / VOC mining from calls, surveys and complaints
- Cross-sell and upsell signal extraction
- Household and group-account consolidation
- Lifetime-value, retention-risk and churn narratives

COMPLIANCE, FINANCE & REPORTING

Audit-grade drafts, not blank pages

- Complaint classification and state-DOI reporting drafts
- Market-conduct exam prep & evidence summarization
- Regulatory filing narratives and disclosure checks
- Reserve, IBNR and claim-trend commentary
- Internal audit trail summarization with source citations

REFERENCE SCORECARD

Fine-Tuned DLM (7B SLM target) vs. Generalist Frontier LLM (zero-shot)

Generalist column: observed zero-shot performance of frontier LLMs (Claude / Azure OpenAI) on representative carrier evaluations. InsightDLM column: design targets for a fine-tuned Qwen/Llama/Mistral 7B base on customer data, served in-VPC. A Tier 2 30B–70B domain LLM pushes generative-task numbers higher still.

INSURANCE TASK / METRIC	GENERALIST FRONTIER LLM (Claude / OpenAI, zero-shot)	INSIGHTDLM 7B FINE-TUNED SLM (Qwen / Llama / Mistral base)
Submission appetite-match accuracy · Top-1	~75%	**≥90% (target)**
ACORD / SOV extraction · Field-level F1	~88%	**≥96% (target)**
UW referral memo quality · UW rating 1–5	~3.6	**≥4.2 (target)**
FNOL peril classification · Top-1	~85%	**≥92% (target)**
Coverage Q&A grounding · Citation precision	~80%	**≥95% (target)**
Claim / 360 summarization · Adjuster / UW rating 1–5	~3.7	**≥4.3 (target)**
Complaint & regulatory drafts · Compliance acceptance	~70% with rework	**≥90% with minor edits**
Median latency, in-workflow assist · p50 / p95	~900 ms / ~3.5 s	**~150 ms / ~600 ms (target)**
Cost per 1K calls · Typical insurance task	~$5 – $30	**~$0.02 – $0.20 (target)**
Data residency / PII exposure	Egress to hosted API · review cycles	**In-VPC or on-prem · no egress**
Auditability & reproducibility	Vendor-controlled model & updates	**Dataset hash · recipe · commit lineage**

PORTFOLIO ROUTING LOGIC

Same table, but as a decision the router actually makes

TIER 1 · SLM DLM

High-volume, bounded, PII-heavy: submission triage, ACORD/SOV extraction, FNOL triage, claim/service summarization, coverage Q&A, complaint classification.

TIER 2 · LLM DLM

Complex multi-doc reasoning and long-form drafting on your data: UW referral memos, 360 narratives, regulatory filings, denial letters that must hold up to legal review.

TIER 3 · FRONTIER (BEDROCK/AZURE)

Broad-reasoning escalations, executive research, exploratory analytics. Claude / OpenAI via Bedrock or Azure OpenAI secure inference — with no model training on your data.

Net effect across the value chain: moving 60–80% of LLM call volume off hosted frontier APIs onto an owned DLM farm typically cuts AI spend by an order of magnitude, drops in-workflow latency 5–10×, eliminates PII egress, and lifts accuracy on the tasks that drive the business — while keeping Claude / OpenAI available, through Bedrock or Azure, for broad reasoning DLMs aren't designed to do.

WHERE THE MODEL FARM RUNS

Four proven deployment patterns

Pattern A · In Your Cloud VPC

MOST COMMON

Tier 1 and Tier 2 DLMs on vLLM / SGLang on managed GPU instances inside your AWS, Azure or GCP VPC. Tier 3 (Claude / OpenAI) via Bedrock or Azure OpenAI in the same account. Training in the same accounts; datasets in S3 / ADLS / GCS.

Pattern B · Hybrid With Frontier Fallback

Owned DLMs serve high-volume in-VPC workloads. The router escalates low-confidence or rare-domain cases to Claude / OpenAI through Bedrock / Azure. One audit log, one cost dashboard.

Pattern C · On-Prem / Air-Gapped

For strict data-residency or on-prem mandates. Fully-private DLMs on on-prem GPU clusters. No egress to public APIs. Tier 3 either disabled or restricted to vetted external workloads.

Pattern D · Edge for Field Ops

Quantized GGUF / AWQ SLMs on adjuster laptops, inspector handhelds and broker workstations — useful for catastrophe response, property inspection and offline submission intake. Same model, prompts and eval suite as cloud.

BENEFITS SUMMARY

What insurers get from owning their DLM model farm

- **Cost predictability** — capacity-based pricing on Tiers 1–2, frontier only where needed.
- **Latency & SLA control** — sub-second responses inside agent and UW workflows.
- **Domain accuracy** — trained on your submissions, policies, claims, calls.
- **Privacy** — PII/PHI never leaves your security boundary.
- **Auditability** — dataset hash → recipe → commit lineage on every model.
- **No vendor lock-in** — open-weight models, open frameworks, swappable frontier providers.
- **Reusability** — one curation, training and eval stack across LOBs and the value chain.
- **Regulatory fit** — aligned with GDPR, HIPAA, SOC 2, PCI-DSS, CCPA programs.

ENGAGE WITH VERTICALSERVE

Stand up your adaptive DLM model farm — on your data, in your environment

A typical engagement begins with a joint architecture review across submission, underwriting, policy servicing, claims, customer 360 and compliance — mapping which workloads belong on Tier 1 SLM DLMs, Tier 2 LLM DLMs and Tier 3 Bedrock/Azure frontier inference. From there, InsightDLM curates your data, trains and evaluates the first DLMs, and stands up the adaptive router and registry inside your environment. You leave owning the models, the data pipeline, the eval harness and the model farm.

EMAIL
contact@verticalserve.com

WEB
verticalserve.com