Enterprise AI Training Services

Custom Fine-Tuned LLMs for On-Premises Deployment

Multi-Language | Multi-Framework | Multi-Domain

Tailored AI Solutions for Regulated Industries

CONFIDENTIAL

Executive Summary

We deliver production-ready, fine-tuned large language models (LLMs) trained on your proprietary codebase, documentation, and domain knowledge — deployed entirely within your infrastructure. Our solutions ensure complete data sovereignty while delivering measurable performance improvements across the full spectrum of enterprise technology stacks.

Our portfolio spans code-centric training (Java, Python, .NET, Go, and beyond), domain knowledge models for specialized industries (engineering, healthcare, legal, finance), and hybrid RAG+fine-tuning architectures that combine deep domain understanding with dynamic document retrieval. Each engagement produces quantifiable results backed by formal evaluation reports.

Project Portfolio: Proven Multi-Domain Expertise

The following case studies demonstrate our ability to deliver fine-tuned AI solutions across fundamentally different domains — from enterprise software engineering to physical infrastructure design. Each project showcases our end-to-end pipeline from data curation through deployment.

Enterprise Java Code Intelligence — *Software Engineering*

Challenge	A large enterprise Java codebase spanning WildFly, Spring Boot, Kafka, and Elasticsearch needed an AI assistant that understood internal architectural patterns, framework-specific conventions, and could generate contextually aware code — without sending proprietary source code to third-party APIs.
Approach	Fine-tuned DeepSeek Coder 6.7B using LoRA on curated Java source files extracted from production repositories. Training data included framework configurations, service implementations, event-driven patterns, and integration code. The model was trained on an RTX 5090 GPU over a 49-hour training cycle.
Tech Stack	DeepSeek Coder 6.7B (base), LoRA/QLoRA fine-tuning, CUDA 13.1 (Blackwell sm_120), GGUF quantization (Q4_K_M, Q5_K_M, Q8, FP16), vLLM serving, LM Studio for local inference
Outcome	**Achieved 0.4092 eval loss. The fine-tuned model demonstrates deep understanding of WildFly subsystem configuration, Spring Boot service patterns, Kafka producer/consumer implementations, and Elasticsearch query construction — knowledge entirely absent from the base model.**

Airport Parking Structure Design Expert — *Civil Engineering & Infrastructure*

Challenge	Airport authorities and engineering firms need rapid access to expert knowledge on parking structure design — spanning FAA Advisory Circulars, ACRP research reports, IBC/ACI structural codes, traffic flow calculations, ADA compliance, and municipal zoning requirements. This knowledge is scattered across hundreds of PDFs, technical standards, and institutional expertise.
Approach	Built a complete training pipeline: automated collection of FAA Advisory Circulars and ACRP reports, PDF text extraction, and generation of domain-specific instruction/response training pairs. Fine-tuned Qwen 2.5 7B Instruct using LoRA. Designed for a hybrid RAG+fine-tuning architecture where the fine-tuned model internalizes reasoning patterns while a retrieval layer grounds answers in specific code sections, manufacturer specs, and project-specific constraints.
Tech Stack	Qwen 2.5 7B Instruct (base), LoRA fine-tuning, automated PDF ingestion pipeline (FAA/ACRP sources), Qdrant vector database for RAG, hybrid dense+BM25 retrieval, nomic-embed-text for local embeddings, full on-premises deployment
Outcome	**Purpose-built AI expert that reasons about ramp grades, turning radii, structural load calculations, fire code compliance, revenue control systems, and ADA accessibility — grounded in authoritative engineering sources. Demonstrates our ability to train models on non-code, domain-expert knowledge from technical document corpora.**

Enterprise RAG + Fine-Tuning Platform — *Enterprise Knowledge Management*

Challenge	Enterprise customers in regulated industries need AI that understands their domain deeply AND can reference their specific documentation, configuration files, runbooks, and architectural decision records at query time — all without data leaving their infrastructure.
Approach	Developed a production-grade RAG platform that layers document retrieval on top of domain-fine-tuned models. The architecture ingests four document types (Java source code, XML configurations, Markdown/wiki documentation, and PDFs), chunks them with domain-aware strategies, and embeds them locally. At query time, the fine-tuned model — which already understands domain vocabulary and patterns — receives retrieved context and generates grounded, traceable answers.
Tech Stack	vLLM serving (production-grade inference), Qdrant vector store (scalable, self-hosted), domain-aware chunking for code vs. prose vs. config files, local embedding models (nomic-embed-text, bge-large), FastAPI wrapper with API key management, rate limiting, and audit logging
Outcome	**A turnkey on-premises platform that combines the deep domain understanding of fine-tuning with the factual grounding and auditability of RAG. Regulated customers gain traceable AI answers with source citations — critical for compliance teams that need to verify where an answer came from.**

What These Projects Demonstrate

Our project portfolio is not limited to a single programming language or technology domain. These case studies illustrate several critical capabilities:

- **Code → Knowledge transfer:** The same fine-tuning pipeline that trains models on Java source code also trains models on civil engineering standards and regulatory documents. The methodology is domain-agnostic.
- **Multi-modal training data:** We handle source code, PDFs, XML configurations, Markdown documentation, and technical specifications. Your training data doesn't need to be just code.
- **Hybrid architectures:** Fine-tuning alone has limits. RAG alone has limits. Our proven RAG+fine-tuning architecture delivers the deep understanding of fine-tuning with the factual grounding and updatability of retrieval — the best of both worlds.
- **Production-grade delivery:** Every project ships with formal Model Evaluation Reports, quantized models for efficient deployment, serving infrastructure, and comprehensive documentation.

Language & Framework Training Capabilities

Our fine-tuning infrastructure supports training across any programming language and framework. Below is a representative overview of the technology ecosystems we serve.

Programming Languages

Language	Enterprise Frameworks	Training Focus Areas
Java	Spring Boot, WildFly, Kafka, Elasticsearch, Hibernate, Jakarta EE	Enterprise patterns, microservices, event-driven architecture, legacy modernization
Python	Django, FastAPI, Flask, Pandas, PySpark, SQLAlchemy, Airflow, dbt	Data pipelines, ML ops, ETL workflows, API development, scientific computing
C# / .NET	ASP.NET Core, Entity Framework, Blazor, Azure SDK, WPF, MAUI	Enterprise web apps, cloud services, desktop applications, legacy migration
Go	Gin, gRPC, Kubernetes client-go, Cobra, Terraform SDK	Cloud-native services, CLI tools, infrastructure automation, high-performance APIs
TypeScript / JS	React, Next.js, Angular, Node.js, Express, NestJS, Vue.js	Full-stack development, component libraries, API services, build tooling
Rust	Actix, Tokio, Axum, Serde, Diesel	Systems programming, high-performance services,

Language	Enterprise Frameworks	Training Focus Areas
		WebAssembly, safety-critical code
C / C++	Qt, Boost, gRPC, CUDA, embedded SDKs	Embedded systems, real-time applications, HPC, firmware development
SQL & Data	PostgreSQL, Oracle, SQL Server, Snowflake, BigQuery, Spark SQL	Query optimization, schema design, data modeling, migration scripts

Infrastructure & DevOps

Domain	Technologies	Training Focus Areas
IaC & Config	Terraform, Ansible, Pulumi, CloudFormation, Helm, Kustomize	Infrastructure patterns, module design, state management, drift detection
CI/CD	GitHub Actions, GitLab CI, Jenkins, ArgoCD, Tekton	Pipeline design, deployment strategies, testing automation, security scanning
Containers	Docker, Kubernetes, OpenShift, ECS/EKS, AKS, GKE	Orchestration patterns, security hardening, resource optimization, migration
Observability	Prometheus, Grafana, ELK/OpenSearch, Datadog, Splunk, Jaeger	Monitoring setup, alert rules, log analysis, distributed tracing, dashboards

Domain-Specific Knowledge Training

Beyond code, we fine-tune models on your organization's proprietary knowledge — internal documentation, regulatory frameworks, standard operating procedures, and domain expertise. This creates AI assistants that understand your business context, not just syntax. Our airport parking structure project demonstrates this capability in a non-software domain.

Regulated Industries

- Healthcare & Life Sciences — HIPAA-compliant coding standards, HL7/FHIR integration patterns, clinical documentation, FDA submission workflows
- Financial Services — SOX compliance patterns, risk modeling frameworks, trading system architecture, regulatory reporting (Basel III/IV, MiFID II)
- Legal — Contract analysis patterns, compliance rule engines, case management systems, document review automation

- Government & Defense — FedRAMP/NIST compliance, ITAR-controlled documentation, C2 systems, IL4/IL5 deployment patterns
- Energy & Utilities — SCADA/OT integration, grid management systems, safety-critical code review, NERC CIP compliance

Engineering & Physical Infrastructure

- Airport & Transportation Infrastructure — FAA Advisory Circulars, ACRP research, structural engineering codes (IBC, ACI 318), traffic flow modeling, ADA compliance
- Parking Structure Design — Ramp grade calculations, turning radii optimization, revenue control systems, fire code compliance, precast/post-tensioned structural systems
- Construction & Architecture — Building codes, zoning regulations, material specifications, project management documentation, RFI/submittal workflows

Cross-Industry Applications

- Internal Documentation & Knowledge Bases — Wikis, runbooks, SOPs, architectural decision records, and tribal knowledge capture
- Legacy Codebase Modernization — Training models on legacy systems (COBOL, mainframe, VB6) alongside target platforms to accelerate migration
- API & Integration Patterns — Custom models trained on your specific API contracts, integration protocols, and data exchange formats
- Security & Compliance — Secure coding standards, vulnerability patterns, audit trail requirements specific to your regulatory environment

Our Delivery Process

1. Discovery & Data Preparation

We work with your engineering and compliance teams to identify the highest-value training data within your organization. This includes source code repositories, documentation, internal wikis, architecture diagrams, code review comments, runbooks, technical standards, PDFs, and regulatory documents. All data remains within your security perimeter.

2. Training & Fine-Tuning

Using state-of-the-art base models (DeepSeek Coder, Qwen 2.5, Code Llama, StarCoder, Mistral, and others), we fine-tune on your curated datasets using LoRA/QLoRA techniques optimized for enterprise-scale training. Our infrastructure supports models from 1B to 34B+ parameters. For knowledge-heavy domains, we combine fine-tuning with RAG architectures for maximum accuracy.

3. Evaluation & Benchmarking

Every model ships with a comprehensive Model Evaluation Report including objective performance metrics. We benchmark against base models and commercial APIs using task-specific evaluation suites designed around your actual use cases — not generic benchmarks.

4. Deployment & Integration

Models deploy to your infrastructure via industry-standard serving frameworks (vLLM, TGI, Ollama, Triton). We provide integration guides for IDE plugins (VS Code, JetBrains), CI/CD pipelines, chat interfaces, and custom application endpoints. RAG-enabled deployments include the full retrieval stack: vector database, embedding pipeline, and document ingestion tools.

Why Choose On-Premises Fine-Tuned Models

Capability	Our Fine-Tuned Models	Cloud API Alternatives
Data Sovereignty	All data stays on-premises. No external API calls. Full control over model weights.	Data leaves your perimeter. Subject to provider's privacy policy and retention.
Compliance Artifacts	Full training provenance, evaluation reports, audit trails for regulators.	Limited visibility into model behavior. Black-box compliance challenges.
Domain Accuracy	Trained on YOUR code, standards, and patterns. Understands your architecture.	Generic training. Requires extensive prompting to approximate your context.

Capability	Our Fine-Tuned Models	Cloud API Alternatives
Cost at Scale	Fixed infrastructure cost. No per-token pricing. Unlimited inference.	Per-token pricing scales linearly. Enterprise usage can cost $50K–500K+/year.
Availability	Runs in your data center or private cloud. No dependency on external service.	Subject to provider outages, rate limits, and deprecation decisions.

Regulatory Compliance & Data Governance

On-premises AI deployment is not just a technical preference — for many organizations, it is a regulatory requirement. Our architecture is designed from the ground up to satisfy the most stringent data protection, privacy, and security frameworks across healthcare, finance, government, and international jurisdictions. Below is how our delivery model aligns with the regulations that matter to your compliance teams.

HIPAA — Health Insurance Portability and Accountability Act

Healthcare organizations deploying AI must ensure that any system processing Protected Health Information (PHI) meets HIPAA's Privacy Rule, Security Rule, and Breach Notification Rule. In January 2025, HHS proposed major updates to the HIPAA Security Rule that eliminate the old "required vs. addressable" distinction — all safeguards are now mandatory for any entity handling electronic PHI (ePHI), including AI systems.

How our on-premises model addresses HIPAA:

- **Zero data exfiltration:** Training and inference occur entirely within your infrastructure. PHI never leaves your security perimeter, eliminating the risks associated with cloud API calls where data transits third-party servers. No Business Associate Agreement (BAA) with an external AI vendor is required for the model itself.
- **Audit trail and provenance:** Every model ships with complete training documentation: data sources, training parameters, evaluation metrics, and version history. This satisfies HIPAA's requirement that AI tools be included in risk analysis and risk management compliance activities.
- **Access controls and encryption:** Our deployment architecture supports role-based access control (RBAC), multi-factor authentication (MFA), TLS 1.3 encryption in transit, AES-256 encryption at rest, and comprehensive audit logging — all standard HIPAA Security Rule requirements.
- **De-identification support:** For training data containing PHI, we implement automated de-identification pipelines following HHS Safe Harbor or Expert Determination methods.

Automated scrubbing ensures no PHI leaks into training logs, error messages, or debugging outputs.

- **HITECH Act alignment:** Our model delivery includes breach notification procedures, minimum necessary data access policies, and documentation supporting the HITECH Act's enhanced enforcement provisions.

GDPR — General Data Protection Regulation (EU)

The EU's GDPR governs any organization that processes personal data of EU residents, regardless of where the organization is headquartered. For AI systems, GDPR imposes specific requirements around lawful basis for processing, data minimization, the right to explanation, and cross-border data transfers. The EU AI Act (effective August 2025) adds further obligations for AI transparency and risk management.

How our on-premises model addresses GDPR:

- **Data sovereignty by design:** On-premises deployment means personal data never crosses jurisdictional boundaries. No Schrems II transfer concerns, no Standard Contractual Clauses required for the AI processing itself. Data stays within the EU (or any required jurisdiction).
- **Data minimization and purpose limitation:** Our training pipelines process only the data necessary for the stated purpose. Training datasets are curated with documented legal basis, and we support data retention and deletion policies — including the right to erasure (Article 17) through model retraining without the deleted data.
- **Transparency and explainability:** Our Model Evaluation Reports provide the documentation needed to satisfy GDPR's transparency obligations (Articles 13–14) and the right to meaningful information about automated decision-making (Article 22). Training data provenance is fully documented.
- **Data Protection Impact Assessment (DPIA) support:** We provide the technical documentation required for your Data Protection Officer to complete a DPIA, including data flow diagrams, risk assessments, and mitigation measures.
- **EU AI Act readiness:** Our documentation practices align with the EU AI Act's requirements for high-risk AI systems, including risk assessments, activity logs, human oversight provisions, and technical documentation. On-premises deployment simplifies compliance with the Act's data governance requirements for training datasets.

NIST, FedRAMP & FISMA — U.S. Federal Security Frameworks

Organizations serving U.S. federal agencies must comply with the Federal Information Security Modernization Act (FISMA), implement NIST SP 800-53 security controls, and — for cloud deployments — obtain FedRAMP Authorization to Operate (ATO). The NIST AI Risk Management Framework (AI RMF) adds AI-specific governance and risk management

requirements, and NIST is developing SP 800-53 Control Overlays specifically for securing AI systems.

How our on-premises model addresses federal requirements:

- **NIST SP 800-53 alignment:** Our deployment architecture supports the full NIST SP 800-53 Rev 5 control catalog, including access control (AC), audit and accountability (AU), configuration management (CM), identification and authentication (IA), system and communications protection (SC), and system and information integrity (SI) control families.
- **NIST AI RMF compliance:** Our delivery process maps to the AI RMF's four core functions: Govern (training data governance, model lifecycle management), Map (risk identification for AI-specific threats), Measure (evaluation benchmarks, bias testing, performance metrics), and Manage (deployment controls, monitoring, incident response).
- **FISMA-ready deployment:** On-premises deployment within a government agency's accredited environment means the AI system inherits the agency's existing ATO boundary. No separate FedRAMP authorization is required for the model, significantly reducing the compliance timeline.
- **Impact Level support:** Our models can be deployed at IL4 (Controlled Unclassified Information) and IL5 (higher sensitivity CUI and National Security Systems) environments because all processing remains within the accreditation boundary.
- **Continuous monitoring:** Our serving infrastructure integrates with standard SIEM platforms (Splunk, ELK/OpenSearch) and supports continuous monitoring requirements mandated by FISMA and OMB directives.

SOX — Sarbanes-Oxley Act

Publicly traded companies and their auditors must ensure that AI systems involved in financial reporting or internal controls meet SOX Section 404 requirements for documentation, testing, and auditability.

- **Auditable AI:** Every model includes complete training provenance, version control, and evaluation reports — creating the audit trail required for SOX compliance. Model behavior is deterministic and reproducible when deployed with fixed configurations.
- **Internal control integration:** On-premises deployment allows your IT controls team to integrate AI systems into existing change management, access control, and segregation of duties frameworks without dependency on external service providers.

Additional Regulatory Frameworks

Regulation / Framework	Scope	Our Compliance Approach
CCPA / CPRA	California consumer privacy rights, including rights related to automated decision-making and profiling	On-premises processing eliminates third-party data sharing. Full documentation supports consumer right-to-know and right-to-delete requests. Aligns with new CCPA automated decision-making regulations.
GLBA	Financial institution safeguards for consumer financial data (Safeguards Rule, Privacy Rule)	Data never leaves the financial institution's perimeter. Encryption, access controls, and audit trails satisfy Safeguards Rule requirements. No third-party vendor risk introduced.
ITAR / EAR	Export controls on defense-related technical data and dual-use technologies	On-premises deployment ensures ITAR-controlled technical data is never processed by foreign nationals or transmitted outside authorized facilities. Model weights remain within controlled environments.
CMMC	Cybersecurity Maturity Model Certification for Department of Defense contractors	Deployment within CMMC-assessed environments preserves certification status. Our architecture supports Level 2 and Level 3 practice requirements for CUI protection.
ISO 27001 / 42001	International standards for information security management (27001) and AI management systems (42001)	Our delivery documentation supports Statement of Applicability requirements. Model lifecycle management aligns with ISO 42001's AI-specific governance controls.
PIPEDA (Canada)	Personal Information Protection and Electronic Documents Act for Canadian organizations	On-premises deployment satisfies Canadian data residency expectations. Training data governance supports PIPEDA's consent and purpose limitation principles.
State AI Laws (CO, IL, TX, NY)	Emerging U.S. state laws requiring bias audits, pre-deployment safety testing, and algorithmic transparency	Our Model Evaluation Reports include bias testing and performance analysis across demographic categories. Full documentation supports transparency and disclosure obligations.

The On-Premises Compliance Advantage

The common thread across every regulation above is this: on-premises AI deployment fundamentally simplifies compliance. When training data never leaves your infrastructure, when model weights are under your physical and logical control, and when inference happens within your accreditation boundary — entire categories of regulatory risk are eliminated. You are not dependent on a third-party vendor's privacy policy, their data retention practices, or their willingness to sign a BAA. You own the model, you control the data, and you can demonstrate compliance to any regulator or auditor with the documentation we provide.

Next Steps

Ready to explore what a custom fine-tuned model can do for your team? We recommend starting with a focused pilot project.

1. **Discovery Call** — 30-minute conversation to understand your technology stack, pain points, and compliance requirements.
2. **Data Assessment** — We evaluate your training data landscape and recommend a high-impact pilot scope.
3. **Pilot Delivery** — A working fine-tuned model with full evaluation report, typically delivered in 2–4 weeks.
4. **Scale & Expand** — Broaden training across additional codebases, languages, and domain knowledge.

Contact us to schedule a discovery call.

All engagements begin with a no-obligation technical assessment.