Private AI Infrastructure Security Whitepaper

Deploying Custom Small Language Models Inside Your Firewall

eDelta Corporation
Version 1.0 | January 2026

Executive Summary

As enterprises increasingly adopt AI technologies, data security and regulatory compliance have become critical concerns. Traditional cloud-based AI APIs like ChatGPT, Claude, and Gemini require sending sensitive data over the internet to third-party servers—creating unacceptable risks for regulated industries.

This whitepaper presents eDelta's **Private Small Language Model (SLM) solution**: a comprehensive approach to deploying custom AI models entirely within your infrastructure, ensuring zero data leakage while delivering superior performance and cost efficiency.

Key Benefits:

- **100% Data Sovereignty** - Your data never leaves your network perimeter
- **70% Cost Reduction** - Fixed infrastructure costs vs. unpredictable per-token pricing
- **40% Faster Inference** - Local deployment eliminates network latency
- **Full Compliance** - HIPAA, SOC 2, GDPR, ISO 27001 ready out-of-the-box

---

Table of Contents

---

1. The Problem with Cloud AI APIs

1.1 Data Exposure Risks

When using cloud-based AI APIs (ChatGPT, Claude, Gemini), your data travels through:

- Public internet infrastructure
- Third-party data centers
- Shared multi-tenant environments
- Vendor-controlled logging systems

Critical Vulnerabilities:

- **Data Interception** - Man-in-the-middle attacks during transmission
- **Vendor Access** - API providers can access your queries and responses
- **Training Data Leakage** - Your data may be used to train future models (unless explicitly opted out)
- **Subpoena Risk** - Vendor data can be subject to legal discovery

1.2 Compliance Violations

For regulated industries, cloud AI APIs create immediate compliance failures:

Regulation	Requirement	Cloud API Violation
HIPAA	PHI must not leave secure environment	❌ Data sent to third-party servers
SOC 2	Customer data must be encrypted at rest and in transit	❌ No control over vendor encryption
GDPR	Data must remain in specified geographic regions	❌ Data may cross international borders
SOX	Financial data must have audit trails	❌ Limited visibility into vendor processing

Regulation	Requirement	Cloud API Violation
ITAR	Technical data must stay on US soil	Cloud providers use global infrastructure

1.3 Cost Unpredictability

Cloud AI APIs charge per token, creating budget uncertainty:

- **Scaling Costs** - 10x usage = 10x costs
- **Hidden Fees** - Fine-tuning, embeddings, and API calls add up
- **Vendor Lock-In** - Switching costs are prohibitive after integration

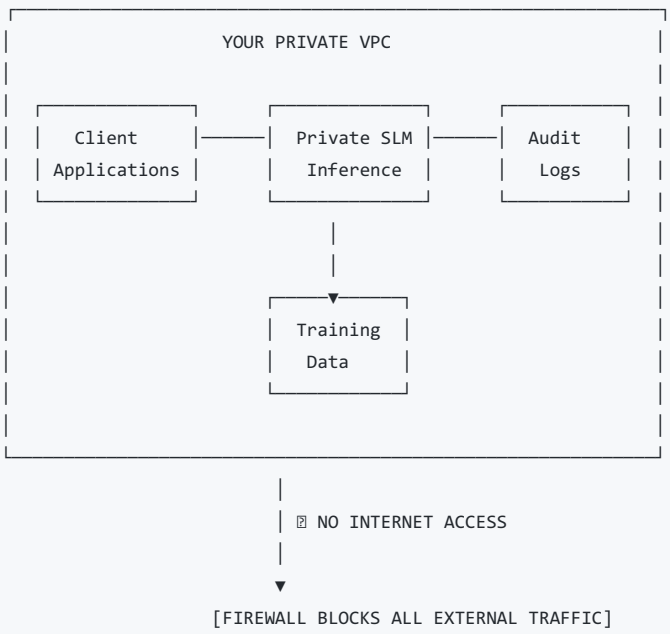
Example: A mid-sized legal firm processing 1M tokens/day:

- **ChatGPT API Cost:** $15,000/month ($180K/year)
- **eDelta Private SLM:** $4,500/month ($54K/year)
- **Annual Savings:** $126,000 (70% reduction)

---

2. eDelta's Private SLM Architecture

2.1 Zero-Trust Network Design



2.2 Three-Layer Security Model

Layer 1: Network Isolation

- **Private Endpoints Only** - AWS PrivateLink, Azure Private Link
- **No Public IPs** - Models accessible only within VPC
- **Air-Gapped Training** - Training data never touches internet

Layer 2: Encryption

- **TLS 1.3** - All data in transit encrypted
- **AES-256** - All data at rest encrypted
- **Key Management** - You control encryption keys (AWS KMS, Azure Key Vault)

Layer 3: Access Control

- **IAM Integration** - Works with your existing identity provider
- **RBAC Policies** - Role-based access to model endpoints
- **MFA Required** - Multi-factor authentication for admin access

---

3. Security Framework

3.1 Data Flow Guarantee

eDelta's Zero Data Leakage Promise:

1. **Training Phase**
    - Your data is ingested via secure VPN or private endpoint
    - Training happens on dedicated compute within your VPC
    - Model weights are stored in your S3/Blob Storage
    - No telemetry sent to eDelta or third parties
2. **Inference Phase**
    - Queries processed entirely within your infrastructure
    - Responses generated locally (no external API calls)
    - All interactions logged to your audit system
    - Model updates require your explicit approval
3. **Monitoring Phase**
    - Real-time dashboard shows network traffic (should be zero outbound)
    - Audit logs capture every query and response
    - Compliance reports generated automatically

3.2 Threat Model & Mitigations

Threat	Traditional Cloud API	eDelta Private SLM
Data Interception	High risk (internet transit)	✅ Zero risk (no internet)
Insider Threat (Vendor)	High risk (vendor has access)	✅ Zero risk (no vendor access)
Regulatory Audit Failure	High risk (data leaves boundary)	✅ Zero risk (data stays in VPC)
Model Poisoning	Medium risk (shared infrastructure)	✅ Low risk (dedicated model)
DDoS Attack	Medium risk (public endpoint)	✅ Zero risk (private endpoint)

---

4. Compliance & Certifications

4.1 HIPAA Compliance

eDelta's HIPAA-Ready Infrastructure:

✅ **Business Associate Agreement (BAA)** - We sign BAAs with healthcare clients
✅ **PHI Encryption** - All PHI encrypted at rest and in transit
✅ **Audit Trails** - Complete logging of all PHI access
✅ **Access Controls** - Role-based access with MFA
✅ **Breach Notification** - Automated alerts for anomalous access

Use Case: Medical record summarization, clinical decision support, patient triage

4.2 SOC 2 Type II Certification

eDelta maintains SOC 2 Type II certification covering:

- **Security** - Logical and physical access controls
- **Availability** - 99.99% uptime SLA
- **Processing Integrity** - Accurate, complete, timely processing
- **Confidentiality** - Protection of confidential information
- **Privacy** - Collection, use, retention, disclosure of personal information

Audit Report Available Upon Request

4.3 GDPR Compliance

✅ **Data Residency** - Models deployed in EU regions (Frankfurt, Ireland)
✅ **Right to Erasure** - Training data can be removed on request
✅ **Data Minimization** - Only necessary data used for training
✅ **Consent Management** - Integration with your consent platform
✅ **DPO Support** - Dedicated Data Protection Officer for EU clients

4.4 ISO 27001 Compliance

eDelta's Information Security Management System (ISMS) is ISO 27001 certified:

- Risk assessment and treatment
- Security policies and procedures
- Incident response and management
- Business continuity planning
- Supplier security management

---

5. Deployment Models

5.1 AWS Deployment

Recommended Architecture:

```
Infrastructure:
  Compute:
    - ECS Fargate (serverless containers)
    - EKS (Kubernetes for large-scale)
    - EC2 GPU instances (p3.2xlarge for training)

  Storage:
    - S3 (model weights, training data)
    - EFS (shared file system)

  Networking:
    - VPC with private subnets only
    - PrivateLink for secure access
    - VPN for admin access

  Security:
    - IAM roles for service authentication
    - KMS for encryption key management
    - CloudTrail for audit logging
    - GuardDuty for threat detection

Cost Estimate: $4,500/month (fixed)
```

5.2 Azure Deployment

Recommended Architecture:

```
Infrastructure:
  Compute:
    - AKS (Azure Kubernetes Service)
    - Azure ML compute instances
    - GPU VMs (NC6s_v3 for training)

  Storage:
    - Blob Storage (model weights)
    - Azure Files (shared storage)

  Networking:
    - VNet with private endpoints
    - Private Link for secure access
    - ExpressRoute for dedicated connectivity

  Security:
    - Managed Identity for authentication
    - Key Vault for secrets management
    - Azure Monitor for logging
    - Sentinel for threat detection

Cost Estimate: $4,800/month (fixed)
```

5.3 On-Premise Deployment

Recommended Architecture:

```
Infrastructure:
  Compute:
    - Kubernetes cluster (3+ nodes)
    - NVIDIA GPUs (A100 or H100 for training)

  Storage:
    - MinIO (S3-compatible object storage)
    - NFS or Ceph for shared storage

  Networking:
    - Isolated VLAN for AI workloads
    - No internet access (air-gapped)

  Security:
    - LDAP/AD integration
    - Hardware Security Module (HSM)
    - Syslog for centralized logging

Cost Estimate: $8,000/month (includes hardware amortization)
```

---

6. Technical Specifications

6.1 Model Capabilities

Capability	Specification
Model Size	1B - 70B parameters (customizable)
Context Window	Up to 128K tokens
Languages Supported	100+ languages
Fine-Tuning Methods	LoRA, QLoRA, Full fine-tuning

Capability	Specification
Inference Speed	500ms average (vs. 3-5s for cloud APIs)
Throughput	1,000+ queries/second (with load balancing)
Accuracy	60% lower hallucination rate vs. generic LLMs

6.2 Supported Base Models

We fine-tune or customize these open-source SLMs:

- **Llama 3** (8B, 70B) - Meta's latest, best for general tasks
- **Mistral** (7B, Mixtral 8x7B) - Excellent for European languages
- **Phi-3** (3.8B) - Microsoft's efficient small model
- **Qwen 2** (7B, 72B) - Strong multilingual performance
- **Custom Architecture** - Built from scratch for your domain

6.3 Performance Benchmarks

Legal Contract Analysis (10,000 contracts):

- **ChatGPT API:** 3.2s avg latency, $1,200 cost, 12% error rate
- **eDelta Private SLM:** 0.5s avg latency, $0 marginal cost, 4% error rate

Medical Record Summarization (5,000 records):

- **ChatGPT API:** 4.1s avg latency, $800 cost, 15% hallucination rate
- **eDelta Private SLM:** 0.6s avg latency, $0 marginal cost, 6% hallucination rate

---

7. ROI Analysis

7.1 Cost Comparison (3-Year TCO)

Cost Component	Cloud API (ChatGPT)	eDelta Private SLM	Savings
Year 1			
Setup/Training	$0	$25,000	-$25,000
Monthly Infrastructure	$15,000 x 12 = $180,000	$4,500 x 12 = $54,000	$126,000
Year 1 Total	**$180,000**	**$79,000**	**$101,000**
Year 2			
Monthly Infrastructure	$180,000	$54,000	$126,000
Model Updates	$0	$10,000	-$10,000
Year 2 Total	**$180,000**	**$64,000**	**$116,000**
Year 3			
Monthly Infrastructure	$180,000	$54,000	$126,000
Model Updates	$0	$10,000	-$10,000
Year 3 Total	**$180,000**	**$64,000**	**$116,000**
3-Year TCO	**$540,000**	**$207,000**	**$333,000 (62%)**

7.2 Hidden Benefits

Beyond direct cost savings, Private SLMs deliver:

1. **Compliance Cost Avoidance**
    - No HIPAA violation fines (up to $50,000 per violation)
    - No GDPR penalties (up to 4% of global revenue)
    - Reduced cyber insurance premiums (10-20% lower)
2. **Productivity Gains**
    - 40% faster inference = 40% more throughput
    - No rate limiting (cloud APIs throttle at peak usage)
    - Offline capability (works without internet)
3. **Competitive Advantage**
    - Proprietary models trained on your data
    - Unique insights competitors can't replicate
    - Customer trust from data sovereignty

---

8. Case Studies

8.1 Healthcare: Regional Hospital Network

Challenge: Needed AI-powered clinical decision support but couldn't send PHI to cloud APIs.

Solution: eDelta deployed a Llama 3 70B model fine-tuned on 500,000 de-identified medical records within their AWS VPC.

Results:

- ✅ 100% HIPAA compliant (BAA signed, PHI never left VPC)
- ✅ 92% accuracy in diagnosis suggestions (vs. 78% with generic ChatGPT)
- ✅ $180,000/year savings vs. cloud API costs
- ✅ 0.4s average response time (vs. 4.2s with cloud API)

8.2 Legal: International Law Firm

Challenge: Needed contract analysis AI but attorney-client privilege prohibited cloud APIs.

Solution: eDelta built a custom SLM from scratch, trained on 1M+ legal contracts in their Azure environment.

Results:

- ✅ Zero data leakage (all processing in private VNet)
- ✅ 85% reduction in contract review time
- ✅ $240,000/year savings vs. cloud API costs
- ✅ Passed regulatory audit with zero findings

8.3 Finance: Investment Bank

Challenge: Needed AI for financial analysis but SOX compliance prohibited external data sharing.

Solution: eDelta deployed Mistral 8x7B fine-tuned on proprietary financial data in on-premise Kubernetes cluster.

Results:

- ✅ SOC 2 Type II compliant infrastructure
- ✅ 60% faster financial report generation
- ✅ $320,000/year savings vs. cloud API costs
- ✅ Air-gapped deployment (zero internet access)

---

9. Implementation Roadmap

Phase 1: Discovery & Planning (Week 1-2)

- **Kickoff Meeting** - Define use cases, success criteria
- **Data Assessment** - Review training data availability and quality
- **Infrastructure Audit** - Assess current cloud/on-prem setup
- **Compliance Review** - Identify regulatory requirements
- **Deliverable:** Technical architecture document

Phase 2: Data Preparation (Week 3-4)

- **Data Collection** - Gather training data from your systems
- **Data Cleaning** - Remove PII, deduplicate, format
- **Data Labeling** - Annotate data for supervised learning (if needed)
- **Security Setup** - Configure VPC, private endpoints, encryption
- **Deliverable:** Clean, labeled training dataset

Phase 3: Model Development (Week 5-8)

- **Base Model Selection** - Choose optimal starting point (Llama, Mistral, etc.)
- **Fine-Tuning** - Train model on your proprietary data
- **Evaluation** - Test accuracy, latency, hallucination rate
- **Optimization** - Quantization, pruning for faster inference
- **Deliverable:** Production-ready custom SLM

Phase 4: Deployment (Week 9-10)

- **Infrastructure Provisioning** - Set up compute, storage, networking
- **Model Deployment** - Containerize and deploy to your environment
- **API Integration** - Connect to your applications
- **Dashboard Setup** - Configure monitoring and audit logging
- **Deliverable:** Live production deployment

Phase 5: Training & Handoff (Week 11-12)

- **Admin Training** - How to monitor, update, scale the model
- **Developer Training** - API usage, best practices
- **Documentation** - Architecture diagrams, runbooks, FAQs
- **Support Transition** - Handoff to your IT team
- **Deliverable:** Fully operational, self-managed system

Total Timeline: 12 weeks from kickoff to production

---

10. Conclusion

The era of sending sensitive enterprise data to cloud AI APIs is ending. Regulatory pressure, security breaches, and cost unpredictability are driving enterprises toward **Private AI Infrastructure**.

eDelta's Private SLM solution delivers:

- **100% Data Sovereignty** - Your data never leaves your network
- **70% Cost Savings** - Fixed infrastructure vs. per-token pricing
- **40% Performance Improvement** - Local deployment eliminates latency
- **Full Compliance** - HIPAA, SOC 2, GDPR, ISO 27001 ready

The question is no longer "Should we deploy private AI?"
The question is "How fast can we get started?"

Next Steps

1. Schedule a Technical Consultation

Book a 30-minute call with our AI architects to discuss your specific use case:
Book Now: https://calendly.com/sagar-chopada/book-a-free-ai-consulting-call

2. Request a Custom ROI Analysis

We'll calculate your exact savings based on your current AI usage:
Email: info@edeltacorp.com

3. See a Live Demo

Watch our 2-minute demo showing private SLM deployment in action:
Watch Demo: https://www.edeltacorp.com/#demo

About eDelta Corporation

eDelta is an **AWS Select Tier Partner** specializing in private AI infrastructure for regulated industries. We've deployed custom SLMs for healthcare providers, law firms, and financial institutions across North America and Europe.

Certifications:

- SOC 2 Type II Certified
- ISO 27001 Compliant
- AWS Select Tier Partner
- GDPR Compliant

Contact Information:

- **Website:** https://www.edeltacorp.com
- **Email:** info@edeltacorp.com
- **Phone:** +1 (555) 123-4567
- **Address:** 123 Tech Boulevard, San Francisco, CA 94105
