PLATFORM ARCHITECTURE

EXECUTIVE SUMMARY

Enterprise On-Premise Agentic AI

Built To Be Your AI Fortress

Version 1.0 | April 2026

www.fortalezaai.com

CONFIDENTIAL

Executive Summary

Fortaleza AI is a complete enterprise Agentic AI platform designed for on-premise deployment in regulated industries. Unlike cloud-based AI solutions that require sending sensitive data to external servers, Fortaleza AI ships as a single Docker Compose stack that runs entirely behind your firewall — from LLM inference to agent orchestration to observability.

This executive summary provides a strategic overview of the platform architecture, security capabilities, deployment models, and business value for CTOs, CISOs, enterprise architects, and engineering leaders evaluating AI solutions where data sovereignty and regulatory compliance are non-negotiable.

The Problem We Solve

Enterprises in healthcare, financial services, manufacturing, legal, insurance, and government face a fundamental tension: they need the productivity gains that AI agents deliver, but their regulatory and security requirements make cloud-based AI platforms difficult or impossible to implement. Every API call to a cloud AI provider creates data exposure risk that compliance teams must accommodate or reject outright.

Fortaleza AI eliminates this trade-off, delivering enterprise-grade Agentic AI capabilities while keeping every byte of data on infrastructure the customer controls.

Key Differentiators

Zero External Data Transmission — No cloud API calls, no telemetry, no data leaving your network boundary unless explicitly configured

Dual-Layer AI Security — LLM Guard (ML-based) combined with NeMo Guardrails (semantic dialog control) for defense-in-depth

Self-Hosted Observability — Langfuse traces every prompt, tool call, and response with audit-ready logs

Open-Source Foundation — Built on LangChain, LangGraph, Ollama, ChromaDB, PostgreSQL — no vendor lock-in, with compatibility with all the major LLMs that you might already be using.

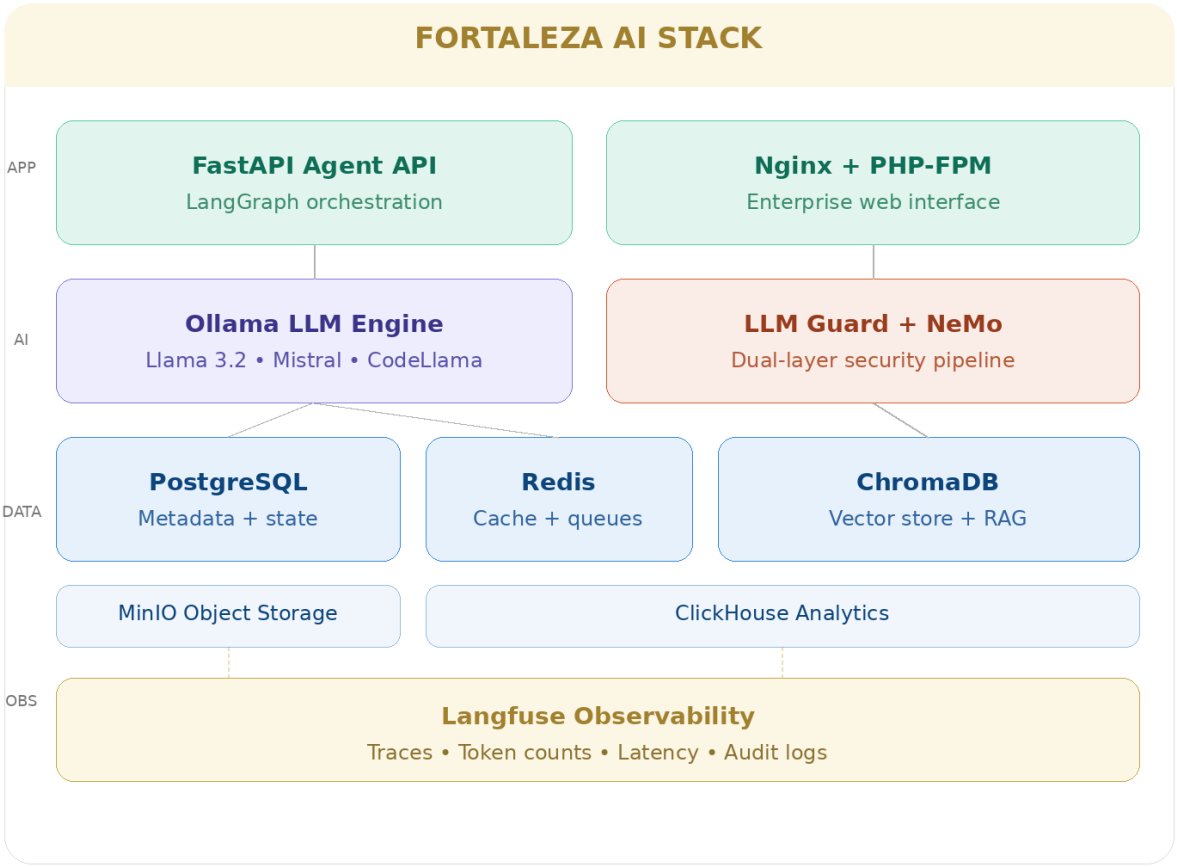
Single-Command Deployment — Docker Compose stack with 16 services, deployable in under 30 minutes, with only specific models downloaded that you configure

80–90% Cost Reduction — Fixed monthly licensing vs. unpredictable per-token API billing

Platform Architecture

The platform comprises containerized services organized into four logical layers: Application, AI and Security, Data, and Observability. Every container runs on customer infrastructure with zero external dependencies.

Architecture Layers



Layer	Components	Responsibility
Application	FastAPI, Nginx, PHP-FPM	REST API, agent orchestration, web interface, API gateway, session management
AI + Security	LangChain, LangGraph, Ollama, LLM Guard, NeMo Guardrails	LLM inference, input/output scanning, PII protection, prompt injection detection
Data	PostgreSQL, Redis, ChromaDB, MinIO	Metadata persistence, caching, object / document storage
Observability	Langfuse, ClickHouse, MinIO	LLM tracing, token usage, latency metrics, cost tracking, audit logging

Agent Orchestration

Fortaleza AI uses LangGraph for multi-agent orchestration, providing a graph-based execution model that supports complex workflows, conditional routing, and stateful conversations. The platform ships a templated agent structure that allows you to choose which LLMs and configurations that you want to use, including:

Ollama (Local LLM): With models like Llama 3.2, Mistral, and CodeLlama. All inference runs locally with zero external API calls — ideal for data-sensitive workloads.

External LLMs (Cloud LLM): Optional integration with OpenAI or other cloud providers for tasks requiring frontier model capabilities on non-sensitive data. Entirely optional and can be disabled.

The LangGraph directed-graph execution model enables conditional tool calling, multi-step workflows with validation, stateful conversation memory, parallel agent execution, and automatic retry with configurable recovery policies.

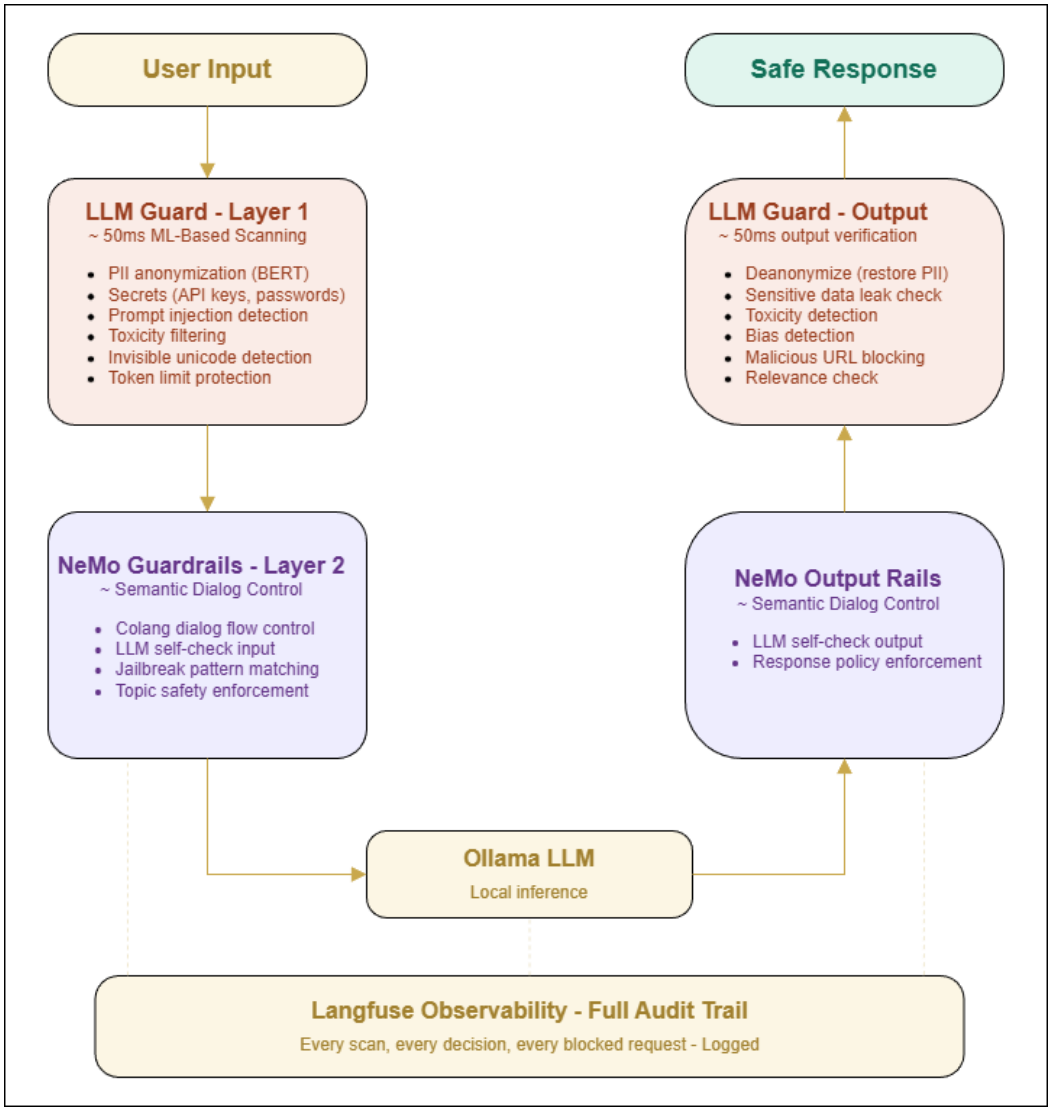
Security Architecture

Security is the architecture, not an add-on. Every user input passes through a dual-layer security pipeline before reaching the LLM, and every LLM response passes through corresponding output rails before delivery. The entire flow is traced and audit-logged.

Dual-Layer Security Pipeline

Layer 1 — LLM Guard (ML-Based): The first line of defense provides fast machine-learning-based scanning. Input scanners handle PII anonymization, API key detection, prompt injection classification, toxicity filtering, and token limit enforcement. Output scanners detect sensitive data leaks, toxicity, bias, and malicious URLs.

Layer 2 — NeMo Guardrails (Semantic Dialog Control): NVIDIA’s NeMo Guardrails uses the LLM itself to perform semantic analysis. Input rails enforce conversation boundaries via Colang dialog flows, jailbreak pattern matching, and topic safety. Output rails verify factual consistency and policy compliance.



Threat Coverage Matrix

Threat Type	LLM Guard	NeMo Guardrails
PII (SSN, credit cards, PHI)	Fast ML detection	—
API keys / passwords / secrets	Pattern + ML detection	—
Prompt injection	ML classifier	LLM semantic check
Jailbreak attempts	Basic detection	Colang pattern library
Toxicity / harmful content	Fast ML scoring	LLM evaluation
Off-topic requests	—	Dialog flow control
Sensitive data leakage	Output leak detection	LLM self-check

Air-Gapped Operation

The entire security pipeline can operate without network connectivity. ML models are pre-downloaded during initialization and stored in shared Docker volumes, ensuring full functionality in air-gapped and classified environments.

Data Sovereignty and Compliance

Data sovereignty is a consequence of the architecture. Because every component runs on customer infrastructure, there is no mechanism by which data could leave the network boundary, even accidentally. LLM inference, vector embeddings, observability, and security scanning all run locally with all external telemetry disabled.

Regulatory Framework Support

HIPAA: Protected health information never leaves the covered entity's network. No BAA required with cloud AI providers.

SOX: Financial data remains within the organization's control boundary. Langfuse audit logs provide required traceability.

GDPR: Personal data processed within EU infrastructure never crosses jurisdictional boundaries.

SOC 2 / FedRAMP: Security architecture aligns with Trust Services Criteria. On-premise deployment satisfies data residency requirements.

Observability and Audit

The platform includes a complete self-hosted observability stack built on Langfuse v3. Every AI interaction — from the initial user prompt through security scanning, agent reasoning, tool calls, and final response — is captured as a structured trace and stored in ClickHouse for high-performance analytics.

Traced data includes: every prompt and response with full content, token counts per request, latency breakdown by component, security guardrail decisions with scan details, agent tool call chains with arguments and return values, and session metadata for multi-tenant environments.

Deployment Models

Fortaleza AI supports multiple deployment models to match enterprise infrastructure strategies. The platform ships as Docker containers that can run on any Linux or Windows (WSL2) host with Docker support.

Deployment Model	Description	Best For
On-Premise Bare Metal	Docker Compose on dedicated servers with automated import scripts	Healthcare, banks, manufacturers with existing data centers
Private Cloud / VPC	Same Docker Compose stack on AWS, Azure, or GCP VMs within an isolated VPC	Organizations using cloud infrastructure with network isolation
Kubernetes	Helm charts with auto-scaling of stateless services	Enterprise-scale deployments requiring horizontal scaling
Air-Gapped	USB/media import with offline licensing	Classified or fully isolated environments

Scaling Tiers

Tier	Infrastructure	Concurrency	Method
SMB	Single server	10–50 users	Docker Compose
Mid-Market	Multi-server	50–200 users	Docker Compose + load balancer
Enterprise	K8s cluster	200+ users	Kubernetes + Helm

Total Cost of Ownership

Fortaleza AI replaces unpredictable usage-based cloud AI billing with fixed monthly licensing, delivering 80–90% savings for typical enterprise workloads.

Cost Category	Cloud AI (Annual)	Fortaleza AI (Annual)
API / Inference costs	$600,000 (usage-based)	Per integration (T&M)
Platform licensing	$0–$60,000	$96,000 ($8K/month, depending on tier)
Infrastructure	$120,000 (cloud compute)	$0 (runs on existing infrastructure)
Implementation	$100,000–$500,000	$15,000–$75,000 (one-time)
Compliance add-ons	$60,000–$180,000	$0 (built-in)
Year 1 Total	$720,000–$960,000	$117,000–$165,000

Cloud AI costs assume 100 employees with moderate usage. Year 2+ Fortaleza costs drop to licensing only ($60K/year) as hardware is amortized.

Open-Source Foundation

Every component in the Fortaleza AI stack is built on proven open-source technology, providing transparency (source code availability), portability (no vendor lock-in), and community-driven improvement.

Component	Project	License	Role
Agent Framework	LangGraph / LangChain	MIT	Multi-agent orchestration, tool calling
LLM Runtime	Ollama	MIT	Local model inference
Security (Layer 1)	LLM Guard	MIT	ML-based input/output scanning
Security (Layer 2)	NeMo Guardrails	Apache 2.0	Semantic dialog control
Vector Database	ChromaDB	Apache 2.0	RAG document search
Observability	Langfuse	MIT (EE)	LLM tracing and audit
Relational Database	PostgreSQL	PostgreSQL	Metadata and state
PII Detection	Microsoft Presidio	MIT	Named entity recognition

Product Roadmap

2026: Foundation and Verticals

Q1–Q2: Core platform v1.0 release with complete Docker Compose stack, export/import distribution system, and enterprise licensing.

Q3: Healthcare vertical with HIPAA compliance validation, clinical documentation agents, and EHR integration templates.

2027: Scale and Ecosystem

Manufacturing vertical with predictive maintenance agents. Financial services vertical with fraud detection agents, risk analysis tools, and core banking integrations. Federated deployment for multi-site enterprises. API marketplace for third-party integrations. Partner certification program.

2028 and Beyond

Industry-specific fine-tuned models. Advanced multi-agent collaboration with shared knowledge bases. Quantum-safe encryption readiness. Visual agent builder for low-code workflow creation.

Getting Started

Fortaleza AI can be deployed to a production environment in under 45 days through a structured engagement:

Week 1 — Infrastructure Assessment: Audit server environment, validate hardware requirements, network configuration, and security policies.

Weeks 2–3 — Pilot Deployment: Working deployment with 1–2 use cases running on your infrastructure for evaluation.

Week 3 — Security Review: Your security team reviews the stack, security pipeline configuration, and audit logging.

Weeks 4–6 — Production Deployment: Full stack deployment with custom agent configurations, RAG document libraries, and security policies.

Week 6 — Training and Handoff: Comprehensive training with documentation, runbooks, and ongoing support channel setup.

Schedule a Technical Deep Dive

jeff@fortalezaai.com
www.fortalezaai.com

Fortaleza AI, LLC

Built To Be Your AI Fortress.

Glossary of Acronyms — Fortaleza AI Architecture Summary

Acronym	Full Term
AI	Artificial Intelligence
API	Application Programming Interface
AWS	Amazon Web Services
BAA	Business Associate Agreement
CISO	Chief Information Security Officer
CTO	Chief Technology Officer
DB	Database
EE	Enterprise Edition
EHR	Electronic Health Record
EU	European Union
FedRAMP	Federal Risk and Authorization Management Program
FPM	FastCGI Process Manager
GCP	Google Cloud Platform
GDPR	General Data Protection Regulation
HIPAA	Health Insurance Portability and Accountability Act
K8s	Kubernetes (abbreviated)
LLC	Limited Liability Company
LLM	Large Language Model
MES	Manufacturing Execution System
MIT	Massachusetts Institute of Technology (license)
ML	Machine Learning
OTEL	OpenTelemetry
PHI	Protected Health Information
PHP	PHP: Hypertext Preprocessor
PII	Personally Identifiable Information
RAG	Retrieval-Augmented Generation
REST	Representational State Transfer
SMB	Small and Medium-sized Business
SOC 2	System and Organization Controls 2
SOX	Sarbanes-Oxley Act
SSN	Social Security Number
TLS	Transport Layer Security
TTL	Time to Live
URL	Uniform Resource Locator
USB	Universal Serial Bus
VM	Virtual Machine
VPC	Virtual Private Cloud
WSL2	Windows Subsystem for Linux 2