

Cortex AI Platform - Service Definition Document

G-Cloud 14 Framework Submission

Service Name: Cortex AI Platform

Supplier: Cortex Labs

Document Version: 1.0

Date: January 29, 202

1. Service Overview

1.1 Service Description

Cortex AI Platform is a secure, modular AI platform designed for UK Government and regulated organizations requiring the productivity benefits of generative AI without compromising data confidentiality. The platform provides access to frontier AI models (from Anthropic, OpenAI, Google, Mistral, Meta and more) through a secure interface, combined with intelligent document processing, semantic search, and governance controls suitable for OFFICIAL and OFFICIAL-SENSITIVE workloads.

Beyond simple AI chat, Cortex delivers automated intelligence through a sophisticated multi-agent system. Built on advanced agentic architecture, the platform coordinates specialized sub-agents that can autonomously conduct complex analysis, generate insights from structured and unstructured data, create visualizations, and produce comprehensive reports, capabilities that would typically require a dedicated data science team. This means organizations gain not just AI assistance, but automated analytical capacity that scales from simple queries to sophisticated intelligence workflows.

The service addresses a critical challenge facing government organizations: teams need AI to maintain productivity and effectiveness, but cannot use commercial platforms due to data sensitivity requirements. Cortex provides enterprise-grade AI capabilities with government-level security, delivered through familiar browser-based interfaces that require no software installation.

1.2 Service Objectives

Primary Objectives: - Enable secure AI access for organizations handling sensitive information - Reduce document analysis time from hours to minutes through automated processing - Provide governance and oversight for organizational AI usage - Eliminate vendor lock-in through multi-model architecture - Deploy operational AI capability in weeks rather than years

Target Users: - Government departments and agencies - Defence and security organizations - Organizations with strict confidentiality requirements

1.3 Key Differentiators

- EU-Bounded Processing: All AI inference occurs within European regions using AWS Bedrock, ensuring OFFICIAL-SENSITIVE compliance
- Multi-Model Access: Choice of Claude, GPT, Llama, Mistral models - no vendor lock-in
- Government-Grade Governance: Comprehensive admin controls, audit logging, spending limits, semantic guardrails
- Rapid Deployment: Operational within one week for standard deployments
- Modular Architecture: Start with basic Chat, add capabilities as requirements evolve

2. Service Components

2.1 Cortex Chat

Purpose: Primary user interface for AI interactions - conversational queries, multi-turn dialogues, and knowledge work.

Capabilities: - Access to multiple frontier AI models via unified interface - Conversation history and management - Collaborative workspaces for team projects - File upload for document analysis (PDF, DOCX, images) - Extended context windows supporting long documents

User Experience: - Familiar chat-style interface requiring no technical training - Keyboard accessible with screen reader support (WCAG 2.2 AA) - Responsive design supporting desktop, tablet, and mobile browsers - Single Sign-On (SSO) integration via SAML 2.0 and OAuth 2.0

Fair Use Allocation: 1,000,000 tokens per user per month

2.2 Cortex Connect

Purpose: Transform organizational knowledge into accessible, searchable intelligence through document ingestion and semantic search.

Capabilities: - Document ingestion with support for PDF, DOCX, TXT, CSV, HTML formats - Automatic chunking with configurable strategies (size, overlap, semantic boundaries) - Vector embedding generation for semantic similarity matching - Semantic search across entire knowledge base - Question-answering over organizational documents - Automated summarization and key point extraction - Citation and source attribution for all AI responses

Technical Features: - Multiple embedding models optimized for different content types - Project-based organization of knowledge bases - Role-based access control for sensitive document collections - Audit trail of all document queries and access

Fair Use Allocation: 10GB document storage per user, 10,000 document embeddings

2.3 Cortex Connect+

Purpose: Advanced AI automation enabling custom workflows, complex analysis, and bespoke intelligence pipelines.

Capabilities: - Custom pipeline building for multi-step workflows - Automated report generation with charts, tables, and visualizations - Web search integration for current information retrieval - Code execution for data analysis and transformation - Database integration for structured data queries - Dashboard creation and business intelligence - Scheduled automation and batch processing

Advanced Features: - Tool calling enabling web search, calculations, data retrieval - Extended thinking and hybrid reasoning for complex problem-solving - Multi-agent coordination for sophisticated workflows - Output schema design for structured data extraction - Test environment for pipeline validation before production deployment

Use Cases: - Automated intelligence briefing generation from multiple sources - Complex data analysis across structured and unstructured information - Regulatory compliance monitoring and reporting - Automated case file preparation from document collections

Fair Use Allocation: 50GB storage per user, unlimited embeddings, 20 GPU hours compute per user per month

2.4 Cortex Control

Purpose: Governance, oversight, and policy enforcement for enterprise AI deployments.

Capabilities:

Usage Monitoring: - Real-time dashboard showing active users, query volumes, cost breakdown - Metrics by user, department, model, and time period - Token consumption tracking with spend forecasting - Feature adoption analytics

Governance Controls: - Semantic guardrails preventing queries on restricted topics - Spending limits by user, department, or organization-wide - Model access restrictions based on role or classification - Multi-factor authentication enforcement

Audit and Compliance: - Comprehensive audit logs: user activity, authentication events, configuration changes - Query history with full conversation content - Data access patterns and document retrieval logs - Exportable audit reports in CSV and JSON formats - Minimum 12-month audit log retention

Administration: - User provisioning and role assignment - SSO configuration and identity provider integration - Project organization and access control - Policy management and enforcement

3. Technical Architecture

3.1 Infrastructure

Cloud Provider: Amazon Web Services (AWS)

Regions: UK (London), EU (Frankfurt, Ireland, Paris, Stockholm) - customer selectable

Deployment Model: Single-tenant SaaS on public cloud infrastructure

Key Components: - Application servers: Auto-scaling EKS instances across multiple availability zones - Database: AWS RDS (PostgreSQL) with synchronous multi-AZ replication - Storage: AWS S3 with server-side encryption and cross-region replication - AI Processing: AWS Bedrock EU inference profiles for model access - Networking: VPC with private subnets, security groups, network isolation

3.2 AI Models and Processing

Model Access via AWS Bedrock EU: - Claude 4.5 family (Opus, Sonnet, Haiku) - GPT models via Bedrock integration - Llama 3 family (8B, 70B, 405B) - Mistral models (7B, Large, Medium) - Additional open-source models as available. Other models available through alternative hyperscalers (eg Google Vertex).

Processing Guarantees: - All inference occurs within EU regions (Frankfurt, Ireland, Zurich, Spain) - No model training or third-party logging on customer data (contractual guarantee) - Zero Operator Access configurable - Query aggregation across shared infrastructure prevents attribution - Models selected per query with automatic failover

Agentic Capabilities: - Tool calling for web search, calculations, data retrieval - Extended thinking for complex reasoning tasks - Multi-turn conversations with context preservation

3.3 Data Storage and Processing

Customer Data Storage: - Primary: Customer-selected region (UK or EU) - Backups: Separate availability zone within same region - Encryption at rest: AES-256 via AWS KMS - Encryption in transit: TLS 1.2 minimum (TLS 1.3 preferred)

Data Residency: - Customers specify storage region during deployment - All processing occurs within selected geographic boundary - No data transfer outside UK/EU regions - Backup and disaster recovery within same regulatory zone

3.4 Security Architecture

Network Security: - VPC isolation with private subnets for application and data tiers - Security groups restricting inter-service communication to required ports - AWS GuardDuty for threat detection and anomaly monitoring - DDoS protection via AWS Shield

Access Control: - Role-based access control (RBAC) for all platform functions - Multi-factor authentication enforcement for administrative accounts - SSO integration via SAML 2.0 and OAuth 2.0 - Session timeouts after 30 minutes inactivity - IP allowlisting available for additional network restrictions

Data Protection: - All data encrypted at rest using AES-256 - Database encryption via AWS RDS encryption - Document storage via S3 server-side encryption - Customer-managed encryption keys available for enhanced control - Automated key rotation enforced

Monitoring and Logging: - AWS CloudWatch for infrastructure and application monitoring - AWS CloudTrail for API activity and audit trails - Application logging for security events and user actions - Automated alerting on suspicious activity patterns - Log aggregation and correlation for threat detection

4. Security and Compliance

4.1 Certifications and Standards

Current Certifications: - Cyber Essentials Plus: Annual certification (verified January 2026) - Founding Team: SC (Security Check) cleared - Staff Screening: BPSS (Baseline Personnel Security Standard) for all staff - AWS Infrastructure: ISO 27001, SOC 2 Type II, ISO 27017/27018 certified

In Progress: - ISO 27001: ISMS operational, formal certification expected Q3 2025

Compliance Framework: - Secure Software Development aligned with ISO 27001 - Vulnerability management complying with CSA CCM v4.0 - Configuration and change management per recognized standards - Protective monitoring per Cloud Security Principles

4.2 Security Testing

Annual Testing: - CREST-approved penetration testing conducted annually - Cyber Essentials Plus external verification - Access control testing and validation

Quarterly Activities: - Vulnerability scanning of infrastructure and applications - Security configuration reviews - Automated dependency scanning for library vulnerabilities

Continuous Monitoring: - AWS GuardDuty threat detection analyzing VPC flows, CloudTrail events, DNS queries - Application security monitoring for authentication failures and anomalous access - Infrastructure monitoring for resource utilization anomalies

4.3 Data Protection

Encryption: - At Rest: AES-256 encryption for all data storage (databases, documents, backups) - In Transit: TLS 1.2 minimum for all network communication - Key Management: AWS KMS with automated rotation; customer-managed keys available

Data Sanitization: - Cryptographic erasure (deletion of encryption keys) for logical data removal - Multi-pass overwriting for database records - AWS-managed physical media destruction following NIST SP 800-88 Rev. 1 - 30-day grace period for data recovery post-contract termination - Verification of complete data removal with audit trail

Backup and Recovery: - Automated daily backups with 30-day retention - Database snapshots every 6 hours - Point-in-time recovery for last 35 days - Backups stored in separate availability zone - Monthly restoration testing to verify integrity

4.4 Protective Monitoring

Threat Detection: - AWS GuardDuty analyzing network traffic, API calls, authentication events - Automated alerts on failed authentication attempts, unusual access patterns, configuration changes - Application-level security event monitoring - Infrastructure anomaly detection

Incident Response: - Critical alerts page on-call engineering team (15-minute response) - Security incidents escalated to Information Security Committee within 1 hour - Automated containment isolating affected resources - Full investigation and remediation plan within 24 hours - Post-incident review within 5 business days

4.5 Vulnerability Management

Assessment: - Annual CREST penetration testing identifies exploitable vulnerabilities - Cyber Essentials Plus validates security controls - Quarterly vulnerability scans assess infrastructure - Automated dependency scanning detects library vulnerabilities - Continuous monitoring alerts on suspicious activity

Patch Management: - Critical security patches deployed within 48 hours after staging validation - High-severity patches within 7 days - Routine patches during monthly maintenance windows - Emergency patches applied immediately for actively exploited vulnerabilities

Threat Intelligence: - NCSC security advisories and alerts - AWS security bulletins for infrastructure vulnerabilities - Vendor notifications from AI model providers - CVE databases and security mailing lists - GitHub Dependabot alerts for dependency vulnerabilities

5. Service Management

5.1 Service Level Agreement

Availability Guarantee: 99.9% platform availability measured monthly

Availability Calculation: - Monthly uptime percentage = (Total minutes - Downtime) / Total minutes × 100 - Downtime: Complete inability to access Chat or Console interfaces - Partial degradation not counted as downtime - Scheduled maintenance windows excluded (announced 7 days in advance)

Service Credits for SLA Breaches: - 99.9% to 99.0%: 10% monthly service credit - 99.0% to 95.0%: 25% monthly service credit - Below 95.0%: 50% monthly service credit - Maximum credit: 50% of monthly subscription fees

SLA Exclusions: - Third-party service failures beyond Cortex control - Customer network or infrastructure issues - Scheduled maintenance with proper notice - Force majeure events - Customer configuration errors - Cyber attacks

5.2 Resilience and Disaster Recovery

Infrastructure Resilience: - Multi-availability zone deployment eliminating single points of failure - Load balancers distributing traffic across multiple application instances - Database synchronous replication across availability zones - AWS S3 storage with 99.999999999% durability - Auto-scaling maintaining performance under demand spikes

Recovery Objectives: - Recovery Time Objective (RTO): 4 hours for full service restoration - Recovery Point Objective (RPO): Maximum 6 hours data loss - Disaster recovery procedures documented and tested quarterly - Runbooks maintained for common failure scenarios

5.3 Support Services

Standard Support (Included): - Email/ticketing: hello@cortex-labs.ai - Web chat during business hours (9am-5pm GMT, Monday-Friday) - Initial response within one business day - Critical issues receive same-day response - Access to knowledge base and documentation

Priority Support (£10/user/month): - 4-hour guaranteed response during business hours - Phone support access - Dedicated support queue - Named technical contact familiar with deployment - Proactive monitoring alerts - Quarterly account reviews

Dedicated Account Management (£2,000/month): - Monthly strategic reviews - Deployment optimization guidance - Feature roadmap input - Priority access to new capabilities - Executive escalation path

Professional Services: - AI Solution Specialist: £1,000/day for training, implementation, optimization - Custom Integration: £2,000/day for bespoke development work - Security Architecture Review: £5,000 fixed price

5.4 Monitoring and Reporting

Real-Time Metrics (via Cortex Control): - Active users and authentication events - Query volumes and token consumption - Cost breakdown by user, department, model - Model

usage distribution - Response times and performance data - Storage consumption - Feature adoption rates - Semantic guardrail violations - Spending against configured limits

Reporting: - Monthly usage summary emailed to administrators - Quarterly business reviews (Priority Support and above) - Exportable metrics in CSV format for external analysis - Historical trending for capacity planning - Custom reports available on request

5.5 Outage Communication

Communication Channels: - Public status dashboard: status.cortex-labs.ai (updated within 5 minutes of issues) - Email alerts to customer administrators (initial notification within 15 minutes) - In-platform notifications during active incidents - Support channel updates for critical incidents - Phone notification for Priority Support customers during major outages

Post-Incident: - Detailed incident reports within 5 business days - Reports include timeline, root cause, impact assessment, remediation steps - Continuous improvement based on incident learnings

6. Implementation and Onboarding

6.1 Getting Started

Standard Deployment Timeline: - Contract signing to service activation: 5 business days - Administrator access to Console: Immediate upon activation - SSO configuration: 1-2 days (customer-dependent) - User provisioning: Same day as SSO completion - Full operational capability: Within 1 week

Onboarding Process:

Day 1-2: Configuration - Administrator accounts created - Cortex Console access provided - Initial governance policies configured - SSO integration setup initiated

Day 3-4: Training - Live webinar for administrators (Console, user management, governance) - Live webinar for end users (Chat interface, document search, workflows) - Recorded sessions provided for reference - Documentation and quick-start guides delivered

Day 5: Go-Live - User accounts provisioned via SSO - Final configuration review - Support contact established - Service declared operational

6.2 Training and Documentation

Online Documentation: - Comprehensive guides covering all platform features - Configuration reference for Console settings - Best practices for AI usage and prompt engineering - Troubleshooting guides and FAQs - Video tutorials demonstrating workflows - Continuously updated with new features

Training Sessions: - Administrator training: Console configuration, governance, user management (2 hours) - End-user training: Chat interface, document search, basic workflows (1 hour) - Connect+ training: Pipeline development, custom workflows (4 hours, separate engagement) - All sessions recorded and available on-demand

Ongoing Support: - Knowledge base with searchable articles - Community best practices (shared with customer permission) - Regular feature update webinars - Office hours for technical questions (Priority Support)

6.3 Professional Services

AI Solution Specialists: - Context engineering and prompt optimization - Use case development and workflow design - Training for power users and administrators - Implementation workshops - Ongoing optimization guidance - Rate: £1,000/day

Custom Integration Services: - Bespoke workflow development - API integrations with existing systems - Advanced pipeline engineering - Custom dashboard creation - Database connectivity setup - Rate: £2,000/day

Fixed-Price Packages: - Connect+ Implementation Package: £8,000 (5 days specialist time) - Security Architecture Review: £5,000 - Priority Onboarding: £3,000 (5-day accelerated deployment)

7. Service Constraints and Dependencies

7.1 System Requirements

Browser Support: - Microsoft Edge (2021 or later) - Firefox (2021 or later) - Chrome (2021 or later) - Safari (2021 or later)

Network Requirements: - Stable internet connection (minimum 5 Mbps recommended) - TLS 1.2 or higher support - JavaScript and cookies enabled - Firewall allowing HTTPS traffic to cortex-labs.ai domain

Display Requirements: - Minimum resolution: 1280×720 - Recommended resolution: 1920×1080 or higher

User Prerequisites: - No client software installation required - Modern web browser with HTML5, CSS3, JavaScript ES6 support - Organizational email address for authentication

7.2 Third-Party Dependencies

Critical Dependencies: - AWS Infrastructure: Bedrock (AI models), RDS (database), S3 (storage), EC2 (compute) - AI Model Providers: Anthropic (Claude), OpenAI (GPT via Bedrock), Meta (Llama), Mistral - Authentication: Customer identity provider for SSO (optional)

Service Continuity: - Multiple model providers ensure availability if one provider experiences issues - Multi-region deployment provides geographic redundancy - Automated failover between availability zones

7.3 Service Limitations

Current Constraints: - Minimum 10 seats per deployment (reduced to 25 seats for government due to support overhead) - EU-bounded processing limits to AWS Bedrock available regions - Computer use capabilities in beta (full production Q2 2025) - Connect+ custom pipelines require 2-4 hours training for effective use - Third-party integrations may require security review for OFFICIAL-SENSITIVE

Not Included: - Connectivity to Public Sector Networks (PSN, HSCN) - customer must provide if required - On-premises deployment (available as separate Sovereign service) -

Post-quantum cryptography (migration planned per NCSC 2035 roadmap) - Mobile native applications (browser-based access available on mobile)

8. Data Management

8.1 Data Export

Self-Service Export (via Cortex Console): - Conversation histories in JSON format with full metadata - Uploaded documents in original formats (PDF, DOCX, etc.) - Knowledge base contents and embeddings - Custom workflows and pipeline configurations - Configuration settings in JSON/YAML - Complete audit logs in CSV and JSON

Bulk Export Tools: - Automated packaging of all organizational data - Structured archives for easy import into alternative systems - Large exports processed asynchronously with email notification - No restrictions on export frequency or volume

Supported Export Formats: - Conversations and metadata: JSON, CSV - Documents: Original formats preserved - Configuration: JSON, YAML - Audit logs: CSV, JSON - Vector embeddings: NumPy arrays, standard vector formats

8.2 End of Contract

Data Extraction Period: - 30-day grace period with full platform access for data export - Self-service export tools available throughout period - Migration support provided at no additional cost - Complex migrations requiring custom scripts: £2,000/day professional services

Data Deletion: - All customer data securely deleted after 30-day grace period - Cryptographic erasure (encryption key deletion) for logical data - AWS-managed physical media destruction per NIST standards - Audit trail documenting complete data removal - Customer receives deletion confirmation certificate

Extended Retention: - Available upon request for compliance purposes - Charged at standard storage rates (£0.15/GB/month) - Maximum retention: 7 years for regulatory compliance

8.3 Data Import

Supported Import Formats: - Documents: PDF, DOCX, TXT, HTML, CSV, XLSX - Data sources: API integration, manual upload, bulk import tools - Configuration: JSON, YAML for automated setup

Import Process: - Documents uploaded via Console interface - Bulk uploads via API for large document collections - Automatic format detection and processing - Validation and error reporting during import - Progress tracking for large imports

9. Pricing Summary

Cortex AI Platform is offered in three service tiers with clear, transparent pricing:

- Chat: £50/user/month (£510 annual with 15% prepayment discount)
- Chat + Connect: £75/user/month (£765 annual with 15% prepayment discount)
- Chat + Connect+: £125/user/month (£1,275 annual with 15% prepayment discount)

Minimum Commitments: - 25 seats minimum - £25,000 annual contract value minimum - 12-month initial term

Volume Discounts: - 100+ users: 10% discount - 500+ users: Custom pricing

Additional Services: - Priority Support: £10/user/month - Dedicated Account Management: £2,000/month - Professional Services: £1,000-2,000/day based on service type

Fair Use Policy: - Chat tier: 1M tokens/user/month - Connect tier: + 10GB storage/user - Connect+ tier: + 50GB storage/user, 20 GPU hours/user/month

For complete pricing details including overage charges, add-ons, and payment terms, refer to the separate Cortex AI Platform - Pricing Document.

10. Ordering and Procurement

10.1 G-Cloud Framework Ordering

Direct Award Process: 1. Review service description and pricing on Digital Marketplace 2. Contact Cortex Labs: hello@cortex-labs.ai 3. Receive detailed proposal and technical documentation 4. Security and compliance review (if required) 5. Issue purchase order via G-Cloud framework 6. Service activation within 5 business days

Further Competition: Available for requirements over £100,000. Cortex Labs participates in competitive tender processes under G-Cloud framework terms.

10.2 Required Information

To expedite procurement, please provide:

- Estimated user count
- Desired service tier (Chat, Connect, or Connect+)
- Contract term preference (monthly vs annual)
- Region preference (UK or EU)
- SSO integration requirements
- Any special security or compliance requirements
- Preferred go-live date

Typical Quote Timeline: 2 business days

10.3 Payment Terms

- Monthly Billing: Net 30 days from invoice date
- Annual Prepayment: Single invoice with 15% discount
- Purchase Orders: Accepted from all UK public sector organizations
- Payment Methods: Bank transfer (BACS/CHAPS)

10.4 Contract Terms

- Initial Term: 12 months from activation
- Renewal: Automatic for successive 12-month periods
- Notice Period: 90 days for non-renewal
- Seat Changes: Add anytime (prorated); reduce at renewal only
- Tier Changes: Upgrade anytime; downgrade at renewal only

11. Contact Information

11.1 Sales and Pre-Sales

Email: hello@cortex-labs.ai

Website: <https://cortex-labs.ai>

Response Time: Typically within one business day

For Quotes and Proposals: - Technical architecture discussions - Security and compliance questions - Pricing and commercial terms - Implementation planning - Proof-of-concept arrangements

11.2 Support

Standard Support: - Email: hello@cortex-labs.ai - Web Chat: Available in Cortex Console (9am-5pm GMT, Mon-Fri) - Knowledge Base: docs.cortex-labs.ai

Priority Support: - Phone support available to Priority Support customers - Dedicated support email with 4-hour SLA - Named technical contact

11.3 Security and Compliance

For security-related inquiries, compliance questions, or to request: - Security architecture documentation - Penetration testing reports - Compliance certifications - Data processing agreements - Security incident reports

Contact: hello@cortex-labs.ai with subject line "SECURITY INQUIRY"

12. Document Information

Document Version: 1.0

Publication Date: January 29, 2026

Next Review Date: June 30, 2026

Document Owner: Cortex Labs Product Team

Revision History: - Version 1.0 (January 29, 2026): Initial publication for G-Cloud 14 submission

Related Documents: - Cortex AI Platform - Pricing Document - Cortex AI Platform - Terms and Conditions - Security Architecture Overview (available under NDA) - Data Processing Agreement (DPA)

This service definition document is provided for informational purposes and forms part of the Cortex Labs G-Cloud 14 service listing. Service specifications, features, and capabilities are subject to change with notice to existing customers. For the most current information, please visit the Digital Marketplace or contact Cortex Labs directly.