Platform Whitepaper

Document Revision April 22, 2025

Contents

1. Introduction

The ConfidentialMind product enables the deployment and use of large language models ("LLM") through enterprise-ready AI endpoints that integrate seamlessly into an organization's IT infrastructure. It gives organizations total control over both the LLM models they deploy and the data these models can access by enabling deployment in on-premises, private cloud, and public cloud environments.

The platform allows organizations to deploy both individual LLM models and complete AI systems such as Retrieval-Augmented Generation (RAG) and Advanced Solution Generation (ASG) systems as ready-to-use endpoints. These endpoints provide standardized APIs that can be directly integrated into existing applications, internal tools, and workflows with minimal development effort.

The product consists of two complementary components:

1. The infrastructure stack that provides GPU resource management, model deployment, and endpoint creation capabilities
2. The AI endpoints layer, which offers ready-to-use complete AI systems with standardized APIs for specific business needs such as document retrieval, complex reasoning, and database agents

The stack and endpoints layer form a comprehensive solution enabling organizations to leverage state-of-the-art LLM technology with their internal confidential data. Both components are licensed together on commercial terms as part of an integrated offering, designed to provide enterprise-grade security, performance, and ease of use.

Figure 1: * ConfidentialMind product overview.

2. AI Endpoints

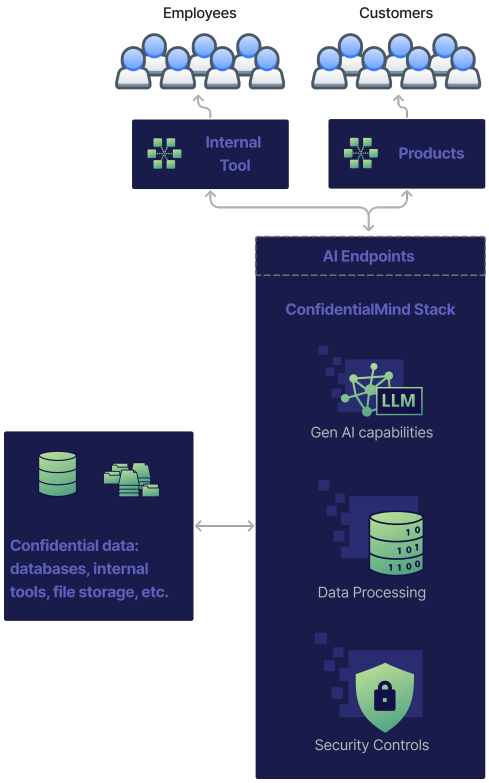

Figure 2: * ConfidentialMind endpoints in use.

ConfidentialMind provides complete AI systems that integrate directly into your products and services through standardized endpoints. When you deploy an endpoint, all required infrastructure components like model deployments, databases, and storage are automatically provisioned and connected.

2.1 Core Endpoint Features

- OpenAI-compatible APIs for drop-in compatibility with existing tools and libraries
- Automatic provisioning of required infrastructure components (models, databases, storage)
- Secure API key authentication and user management
- Configurable data source connections (AWS S3, Azure Blob Storage, and more)
- Standardized endpoints for different AI use cases (RAG, Complex Reasoning, Database Agents)
- Support for both on-premises LLMs and external model providers when needed*
- Streaming responses for better user experience with longer outputs
- Comprehensive documentation and integration examples
- Easy configuration through the administrator interface

2.2 RAG Endpoint

The RAG (Retrieval-Augmented Generation) Endpoint enhances LLM responses with relevant context from your organization's documents. This system provides a standardized API for document ingestion, context retrieval, and augmented generation.

- Automated document processing with intelligent chunking
- Vector storage using PostgreSQL with pgvector
- Multiple retrieval methods (semantic similarity, keyword matching, BM25)
- Flexible filtering by metadata, document groups, or content characteristics
- Document management API for uploading, organizing, and retrieving documents
- Context retrieval API for finding relevant document passages
- Chat completions API for generating responses enhanced with document context

2.3 MCP Agent Endpoint

The MCP (Model Context Protocol) Agent provides a standardized framework for connecting LLMs with external tools and data sources. This extensible system enables natural language interactions with virtually any resource through a consistent protocol, allowing organizations to easily build custom AI agents that can access, analyze, and act on their proprietary systems and data.

- Built on the open Model Context Protocol (MCP) for standardized AI agent development
- Modular architecture with pre-built MCP tools for common data sources
- Extensible framework to develop custom MCP tools for proprietary systems
- Secure protocol with fine-grained permission controls for each tool
- Multi-step reasoning capabilities with intelligent planning and replanning
- Built-in conversation history and context management
- Containerized microservice architecture for scalability and isolation
- Development SDK for creating custom MCP tool implementations

The system includes several pre-built MCP tools for common use cases (PostgreSQL databases, document repositories, knowledge bases), while the extensible architecture allows organizations to develop custom tools for their unique systems. The standardized MCP protocol ensures all tools work consistently with the agent framework, regardless of whether they're pre-built or custom implementations.

**It is possible to make use of other LLMs outside the ConfidentialMind product. For example, OpenAI GPT-4o hosted by Azure OpenAI or AWS Bedrock models. In this case some data will be transferred outside of the ConfidentialMind product. Selecting a trustworthy external LLM is recommended when their use is required.*

3. Infrastructure Stack

The ConfidentialMind infrastructure stack is a Kubernetes-based platform optimized for deploying and managing LLMs and AI endpoints on any infrastructure that supports Kubernetes and has suitable GPU resources for LLM computation.

The core of the stack combines industry-standard Kubernetes components with proprietary ConfidentialMind technology for seamless operation. Our proprietary IP focuses on critical enterprise needs: data security, authentication, GPU resource management, model deployment orchestration, and endpoint provisioning.

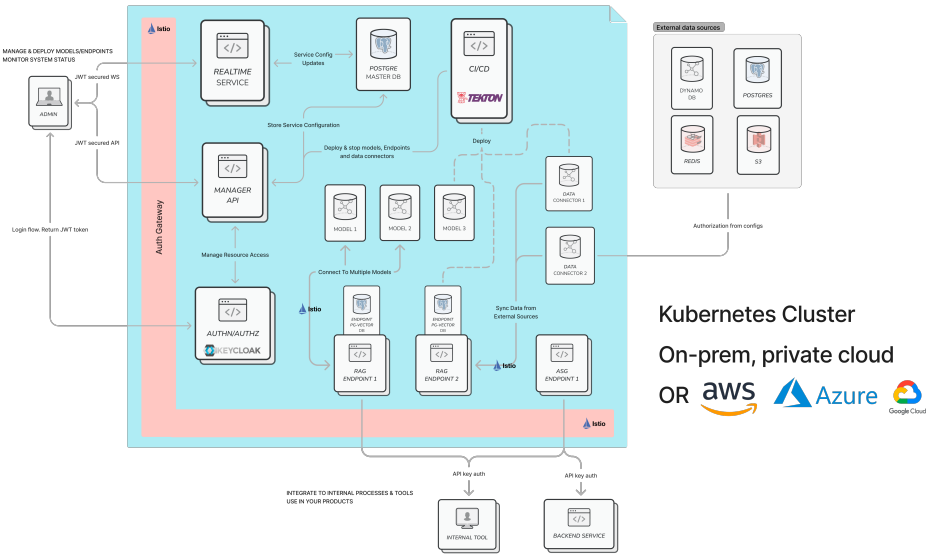

Figure 3: Stack architecture. More information in Appendix 1.

The stack can run on any infrastructure that can run Kubernetes and has the necessary compute capabilities for LLMs. This means that the solution can be deployed on on-premises servers, private clouds, or public cloud environments. Air-gapped installations are also supported,

enabling fully isolated offline deployments. We also support the following cloud infrastructure vendors: Microsoft Azure, Google Cloud Platform, and AWS.

We continuously enhance the stack with new endpoint types, model support, and data connectors. These connectors facilitate secure and efficient data exchange between AI endpoints and various data sources such as SQL databases, document repositories, and knowledge management systems.

The overall purpose of the ConfidentialMind stack is to provide the foundation for rapid deployment of AI endpoints that can securely work with an organization's most confidential data. This is achieved through:

1. Enterprise-grade infrastructure with comprehensive data connectivity and security controls
2. Intuitive portal interface for GPU resource management and endpoint deployment
3. Standardized APIs and abstractions for consistent integration patterns across different AI capabilities

3.1 Stack core features

Read more about technical features and requirements of the Stack from Appendix 1.

- GPU resource management dashboard for monitoring and optimizing model deployments
- One-click deployment and configuration of AI endpoints
- Comprehensive model deployment interface with fine-grained control over GPU allocation
- Automatic provisioning of vector databases (PostgreSQL with pgvector) for RAG endpoints
- Simple ArgoCD-based installation process and installation scripts
- Internal CI/CD for endpoint deployment and updates
- Comprehensive authentication and authorization system
- Auto-scaling of services based on workload demands
- Extensive model support
  - Models from Hugging Face model hub
  - External models from Azure OpenAI, OpenAI API, and AWS Bedrock
- Advanced data processing capabilities including automated document indexing to RAG endpoints

4. Why ConfidentialMind?

Integrating AI systems with confidential organizational data presents significant challenges with traditional approaches. Building these systems typically requires managing complex infrastructure, securing data pipelines, deploying models, and developing integration points—all of which demand specialized expertise and substantial time investments. ConfidentialMind simplifies this process by providing ready-to-use AI endpoints that can be deployed in your secure environment, offering public cloud-like capabilities with enterprise-grade security and control.

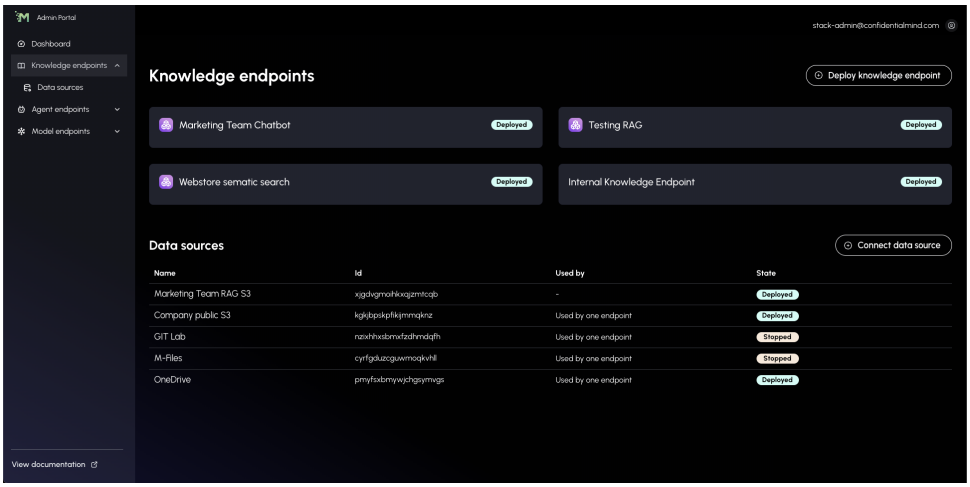

Figure 4: ConfidentialMind management portal.

The ConfidentialMind management portal provides a comprehensive interface for deploying and managing AI endpoints and the underlying infrastructure. This intuitive dashboard gives administrators complete visibility and control over GPU resources, model deployments, endpoint configurations, data connectors, and security policies—all from a single unified interface.

4.1 Comparison with traditional approaches

	With ConfidentialMind	**Without**
Endpoint Deployment	Deploy complete AI systems with a few clicks through the management portal. AI endpoints automatically provision and configure all required components (models, databases, storage).	Build each component of your AI system separately, manage their dependencies, and create custom integration code between components.
Authentication	Configure identity providers or federated login through the management portal. All endpoints automatically inherit the authentication system, with API key management for secure integration.	Implement separate authentication systems for each component, manage API keys manually, and build custom integration between auth systems and AI services.
Model Management	Unified GPU resource dashboard shows model allocation and performance. Deploy models with optimized settings for your hardware. Easily switch models used by endpoints without changing integration code.	Manually manage GPU resources, write custom model deployment code, and rebuild integrations when changing models.
Data Security	Comprehensive security controls with service-level access management and mesh-level security policies. Data remains within your infrastructure with fine-grained permission controls.	Build custom security systems to control access to models and data. Implement your own permission systems and API security controls.
Integration	Standard OpenAI-compatible APIs across all endpoints. Comprehensive SDK and code examples for rapid integration. Streaming support and consistent error handling across all endpoints.	Learn different APIs for each model provider. Build custom integration code for each model and data source. Handle inconsistencies between different vendor APIs.
Enterprise Readiness	Production-grade infrastructure with high availability, monitoring, and scalability built-in. Complete deployment history and rollback capabilities.	Build your own production infrastructure for AI systems. Implement custom monitoring, logging, and scaling capabilities.

Table 1: Comparison of AI integration experience with and without ConfidentialMind

5. Licensing

Both the infrastructure stack and AI endpoints are licensed together on an annual subscription basis. The core components of the platform have proprietary closed-source licensing, while some endpoint templates and example applications are provided under Apache 2.0 or similar open-source licenses to allow for customization and extension.

The licensing model consists of two primary components:

1. Annual enterprise subscription fee paid once per year by each customer
2. Usage-based component calculated based on total VRAM allocated to the cluster

The cluster control plane and management components do not count against the compute units for licensing purposes. This model provides flexibility to scale your AI endpoint deployments based on your specific needs, from small proof-of-concept installations to enterprise-wide production deployments. Detailed pricing information and custom licensing options are available upon request.

6. Contact

If you have any questions, you can reach our technical and sales teams at info@confidentialmind.com.

- For specific technical questions, please contact Mr. Severi Tikkala, CTO at severi@confidentialmind.com and
- For specific sales or business questions please contact Mr. Markku Räsänen, CEO at markku@confidentialmind.com.

ConfidentialMind is made in Espoo, Finland by infrastructure and AI nerds.

7. Appendix

.1 Stack technical specifications

- Easy installation process
  - Simple installation tool to install core ArgoCD and Kubernetes components
  - ArgoCD based installation process for the stack and dependencies
  - ArgoCD based automatic update process for the stack and dependencies
  - Possibility for manual updates
- Core Kubernetes components
  - Istio
    - Ingress gateway
      - All traffic to the cluster is routed through the gateway
      - JWT based authorization with Keycloak
      - Option to enable API-key based access rules for external traffic
    - Service mesh
      - Zero-trust approach: Deployed containers are not allowed to communicate with each other by default (AuthorizationRules)
    - Reverse proxy with Istio envoy filter
      - Provides authentication for workloads that don't implement their own authentication flow
  - Keycloak
    - Identity and access management
    - Integration to organization's existing users and roles
    - Deployed resources and workloads are Keycloak resources. Keycloak provides authorization to manage these resources
  - Tekton pipelines with Kaniko to build application images
    - Pipelines integrate to customer's git repositories enabling automatic deployments from webhook triggers
    - All customer code is built and deployed inside stack, not on ConfidentialMind's servers
  - TLS certs
    - Letsencrypt for automatic TLS generation and renewal
    - Option to use organization's own certificates
  - Monitoring
    - Prometheus configured for monitoring workloads and resources
    - Possibility to enable Istio network audit logs
- Core services
  - LLM inference service
    - Supports most models from Hugging Face and most common LLM architectures
    - Support for NVIDIA, AMD and Intel GPUs, most common CPUs also supported for smaller model inference
    - OpenAI compatible interface
  - External LLM connectors for Azure OpenAI and AWS Bedrock
    - OpenAI compatible interface
  - Microservices
    - Deploy any containerized workload with public or private (authenticated) access
    - Scaling and replica sets automatically available
  - Databases and storage

    - PostgreSQL clusters available on microservice level with vector and graph store support
    - Persistent file storage provisioning from node disks
  - AI services
    - Agent engine (**coming soon**)
    - Automatic indexing of data sources
  - Manager service
    - Protected endpoint for Admin users to manage stack resources such as LLMs, applications and authorization rules. Also accessible through API for more in-depth integration to existing tooling
- Container registries
  - ConfidentialMind secure registry
    - Helm charts for the stack, microservices and inference servers
    - Template application images
  - Bring your own registry
    - Connect your own container registry to easily deploy images from them on the stack

.2 Stack technical requirements

- Kubernetes cluster
  - Managed Kubernetes cluster
  - Kubernetes cluster created using our cluster creation documentation
- Node and node pool requirements. Cluster (worker nodes) autoscaling currently available only cloud.
  - Control plane, 3 nodes with minimum 4-8 vCPUs and 16-32GB of RAM each. Excluded from license pricing
  - Worker nodes, minimum 1 node with recommended setup of
    - +64 CPU cores
    - +80GB of VRAM, for example 1 x NVIDIA A100 or 2 x NVIDIA L40S GPUs
- Network
  - A-record configuration for your custom domain or DNS and certificate setup according to your own infrastructure
  - Information if your network implements load balancing or does the ConfidentialMind stack enable its own load balancer

