ConfidentialMind

1. Introduction

The ConfidentialMind product enables the deployment and use of large language models (“LLM”), and AI enabled applications in an organization’s IT infrastructure. It gives organizations total control over the use of these models as well as the data they have access to by enabling the deployment of LLM and AI enabled applications in on-premises, private cloud, and public cloud environments.

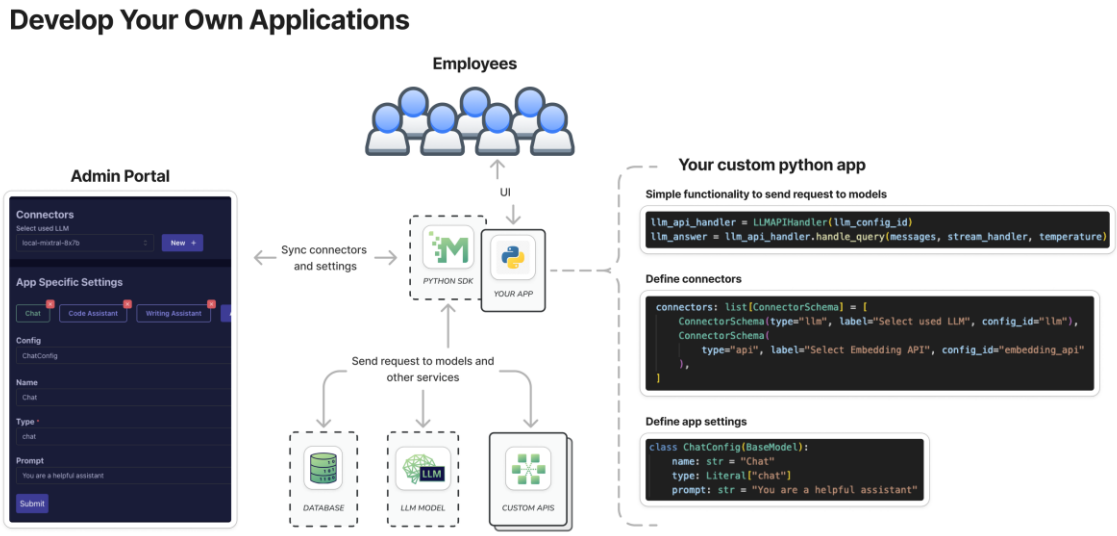
The product consists of two separate parts: (1) the stack that is the infrastructure needed to host and integrate LLMs as well as (2) the application layer, which is a set of applications and templates on top of the stack for various business use cases. The application layer is also where all customer applications will be deployed.

The stack and application layer form the backend and front-end parts of the complete deployment enabling the organization’s employees to use LLMs and business specific LLM-enabled applications with the organization’s internal confidential data. Both stack and the application layer are licensed together on commercial terms as part of a common offering.

Document revision 13.8.2024

2. Workloads

ConfidentialMind stack allows you to run any containerized applications and easily utilize generative AI capabilities such as inference, vector and graph databases from the stack via a graphical user interface and SDK. It enables you to develop and deploy internal applications and tooling, enabled by generative AI, without the fear of data leaks. ConfidentialMind provides its customers example applications and templates licensed under Apache-2.0 license to get started quickly. These examples showcase how to quickly build applications and workloads that use other resources deployed on the stack.



Application layer core features

- Deploy any containerized application or API with one-click deployment
- Create public or private endpoints, or create API-key enabled API-endpoints
- Retrieval augmented generation (“RAG”) resources such as vector database and file storage are fully integrated into the development workflow and come with the infrastructure
- Collection of business area and task specific applications utilizing open source LLMs inside ConfidentialMind stack or your own LLM infrastructure*****
- Iterate faster with the latest technology and innovate internal applications specific to your business
- Example applications and API implementations available
- Built-in extensive generative AI enabled documentation system to support development with ConfidentialMind and answer the most frequent questions fast

- Applications can be configured from the administrator interface with ease

***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. Stack

The ConfidentialMind stack is a Kubernetes-based containerized set of services and infrastructure required to run LLMs and LLM-enabled applications on any infrastructure that supports Kubernetes and LLM-compute.

The core of the stack is comprised of Kubernetes *de facto* standard components configured to work seamlessly together via proprietary ConfidentialMind IP. In addition, the key areas of data security, authentication, ease of software development and AI agent development and management are managed by proprietary ConfidentialMind IP.

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 also be enabled to run on on-premises servers and infrastructure. Air-gapped installations are also possible, enabling fully isolated offline installations. We also currently support the following cloud infrastructure vendors as places to deploy ConfidentialMind: Microsoft Azure and AWS. Other cloud vendors such as GCP and IBM Cloud are coming soon.

We are constantly updating the stack with new AI capabilities and connectors. Connectors are code templates and helper functions to retrieve data from or store data to different data sources such as SQL or document databases.

The overall purpose of the ConfidentialMind stack is to serve as the required infrastructure for the internal developers of an organization to develop LLM-enabled applications on top of the organization's most confidential data. This is achieved through (1) the specific focus in the ConfidentialMind stack on data connectivity and data security, as well as (2) the ease of use of the infrastructure management achieved through the ConfidentialMind portal, and (3) the overall ease of software development on top of the stack achieved through purpose-built industry-leading

abstractions of services and components required for fast and efficient LLM-application development.

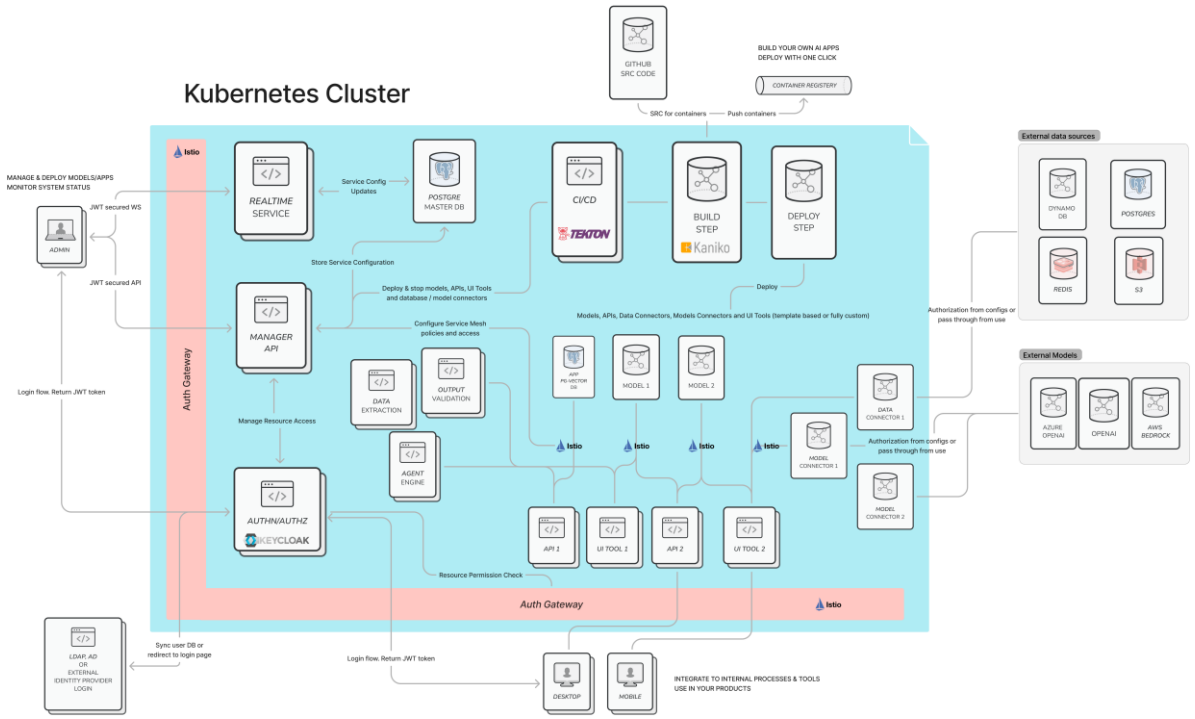


Figure 1 Stack architecture. More information in Appendix 1.

Stack core features

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

- Simple graphical admin user interface for managing the stack and its functionality
- Ease of deployment and management of LLMs
- Ease of deployment and management of microservices that use the LLMs
- Existing backend and frontend application examples that utilize the LLMs from the stack
- Ease of deployment for application databases (currently: Postgres) and vector databases (currently: pgvector)
- Ease of deployment for S3 buckets
- Simple Helm-based installation
- Internal CI/CD for microservices
- Authorization and access management
- Auto scaling of all services
- Model support
    - Models from Hugging Face
    - External models from Azure, OpenAI and AWS bedrock
- Abstracted AI capabilities such as automated data indexing to knowledge graph

4. Why ConfidentialMind?

LLM application development with confidential data is difficult using the currently available tools. Often the deployment environments for confidential data need to be made bespoke, which is both slow and costly. ConfidentialMind makes it easy for you to develop cutting edge LLM applications faster and cheaper and in a more secure way. You get public cloud like capabilities in your own environment.

In addition to all the features in the ConfidentialMind stack, the product also offers an easy-to-use portal interface for controlling the underlying infrastructure irrespective of its location. This includes the full ability to control stack sub-services such as LLMs, data connectors, object storage buckets, databases, and other related components.

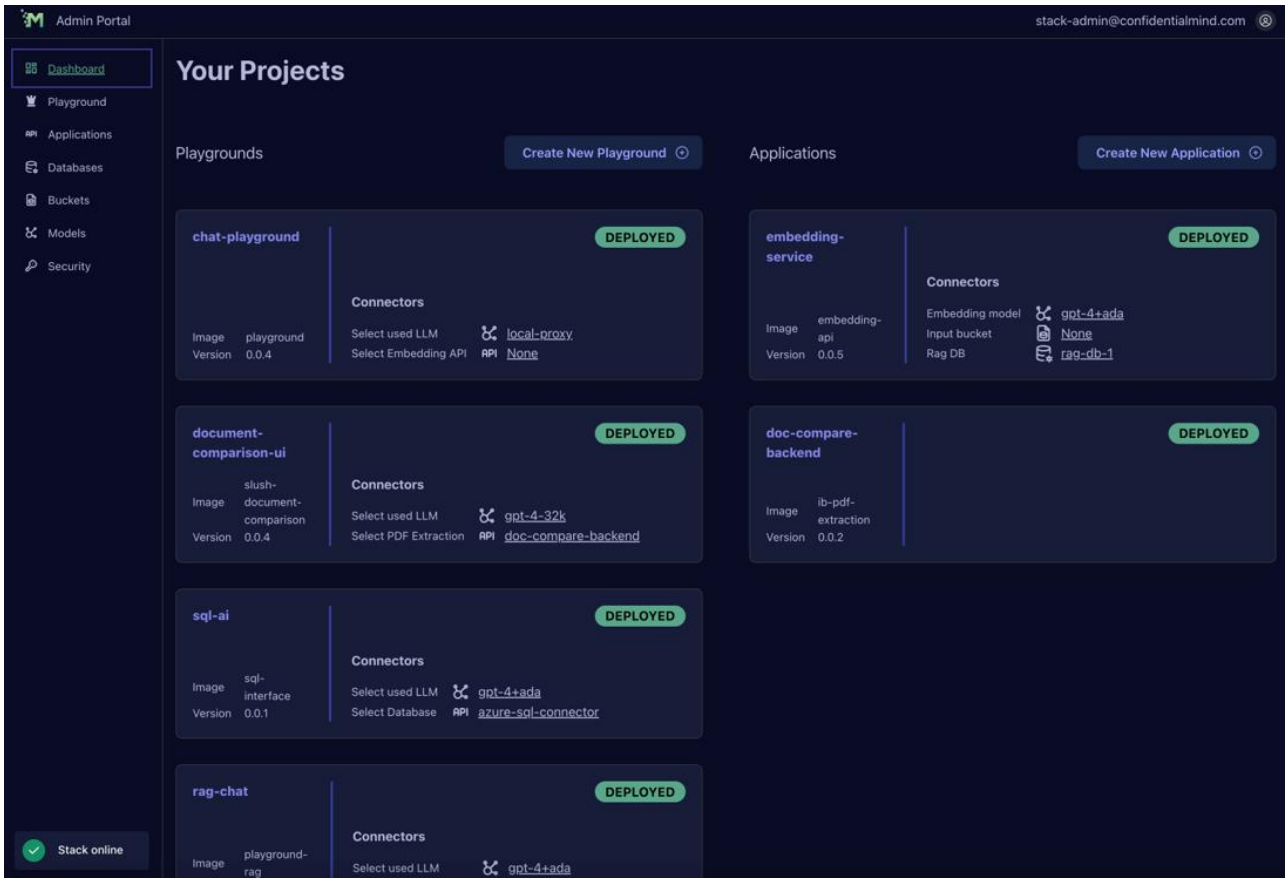


Figure 2 ConfidentialMind management portal.

	With ConfidentialMind	Without
Authentication	The stack provides easy ways to configure almost any identity provider or federated login. The admin portal works out of the box with the authentication you have configured, and all custom services are automatically behind auth proxy.	You need to select and deploy the authentication system yourself and do the FE integration and possible proxy implementation
Authorization	The stack provides easy service level group/user access management through the admin portal. Also mesh level security policies are automatically created based on connected services.	You would need to make a resource or role-based authorization system to make sure different teams/individuals have access to the right services. You would also need to make some system that verifies the stack internal security so rogue developers or other bad actors don't get access to data they should not have access to.
Prototyping to production + CI/CD	The stack works both as a development platform and production ready deployment system, with built-in CI/CD. We are currently working on the support for automatic CI/CD for 3rd party services (connecting your own repository to the stack). The stack will automatically build new images from the repo and services can be updated either automatically or from admin panel.	You need to configure your development environment, production environment and your own CI/CD.
Template code and python SDK	The stack provides python templates for common use cases and an SDK that allows easy communication between services in the stack. For any UI-based templates (Streamlit, etc.) the stack supports automatically adding a reverse auth proxy in from of the service so developers can focus on business logic	You need to start the development from zero and investigate how to combine different services, use the API endpoints of different LLM models, etc.
Configurability	The stack provides an easy way for services to announce configuration options that are automatically configurable through the admin UI. Though this system you can for example test different prompts easily.	For rapid LLM app development you need some way to test different prompts and configurations. Note that the developer of the app might not be the best person to come up with for example the best prompts, so it is good to have a simple way for the domain experts to also test different prompts without code changes.
Existing services in the stack	The stack provides multiple different pre-configured services that are useful for kickstarting development. Through the admin portal a developer can spin up a vector database or file storage bucket with one click and connect them to the service that is being developed. When the application goes into production the same system can be used to connect the app easily to production level vector database or existing file storage without any code changes	The developer would need to create a test environment where these services are somehow available during development. And the production environment would need its own configuration.

5. Licensing

Both stack and the application layer are licensed together on a annual fee basis. The source code for the core components for both are closed, but some application layer components come with Apache 2.0 or similar open-source license allowing for follow-on development.

Currently, the licensing model has two main elements:

(1) Annual enterprise license paid once annually by each customer
(2) Variable element based on compute units (CU) each CU is either 32 CPU cores or 48GB of GPU VRAM

Note that the cluster control plane does not count against the CUs for license fee purposes. Pricing information is readily 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.

Appendix 1. Stack technical features and requirements

Stack technical specifications

- Easy installation process
    - Terraform and ansible based install script to create cluster and configure the values for the helm-chart
    - Helm chart-based configuration and installation
    - Operating System image with preconfigured Kubernetes cluster and pre-installed stack (**coming soon**)

- 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 and Grafana configured to 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
        - S3 buckets (experimental)
        - 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

Stack technical requirements

- Kubernetes cluster
    - Managed Kubernetes cluster
    - Kubernetes cluster created by ConfidentialMind OS images (**coming soon**)
- 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