Varys AI Documentation

# 1. Overview

Varys AI is an on-premises Large Language Model (LLM) platform designed for businesses that need secure, private, and customizable AI capabilities without sending their data to external providers like ChatGPT, Gemini, or Claude. Unlike SaaS-based AI tools, Varys AI runs within your controlled environment (on-premise or GPU infrastructure), ensuring data confidentiality, compliance, and adaptability to business workflows.

# 2. Problem Statement

## The Need for Secure and Private AI Solutions

Businesses are rapidly adopting AI for knowledge management, decision-making, and automation. However, existing AI solutions pose significant challenges:

* Data Privacy Risks: Using third-party models like ChatGPT, Gemini, or Claude means sensitive data leaves your environment, creating risks of data leaks, misuse, and non-compliance with regulations (GDPR, HIPAA, etc.).
* Lack of Control Over Models: Third-party AI APIs cannot be customized extensively, leaving businesses dependent on generic models.
* Vendor Dependency: Relying on external models for all queries creates ongoing dependency and limited flexibility.
* Integration Limitations: Current solutions don’t easily integrate with internal systems like file servers and proprietary datasets.
* Unpredictable Costs & Latency: External API usage comes with high recurring costs and latency, not ideal for secure operations.

## Core Problem Varys AI Solves:

* Run AI locally on GPU servers you own or lease, without sending data to third-party APIs.
* Ingest and query internal files and proprietary knowledge securely.
* Maintain full control over customization, data handling, and performance.
* Enable project-based collaboration and contextual AI interactions.

# 3. Key Features (MVP)

* Secure Chat Interface: Conversational AI powered by your uploaded documents.
* Vault System: Centralized file storage for knowledge ingestion.
* Projects & Contextual AI: Group files and conversations under specific projects.
* File Ingestion with Chunking & Embeddings: Handle PDFs, Word docs, text files, etc.
* Local Model Integration via MCP: Connect to open-source models running on your GPU infrastructure.
* Full Privacy & Compliance: No data leaves your environment.

# 4. Target Users

* Enterprises & SMEs that want full control over their AI stack.
* Financial & Legal Firms handling confidential information.
* Healthcare Organizations requiring strict compliance.
* Manufacturing & Engineering Companies with proprietary documents.

# 5. Future Scope

## 5.1 Advanced Model Capabilities

* Fine-Tuning for Specific Domains: Build custom AI models for law, healthcare, finance, or manufacturing.
* Multi-Model Orchestration: Switch between local open-source LLMs and secure external APIs when needed.
* Autonomous AI Agents: Automate tasks like document summarization, report creation, and knowledge extraction.

## 5.2 Enhanced Security & Governance

* Role-Based Access Control (RBAC): Manage user permissions effectively.
* Detailed Audit Logs & Usage Analytics: Track every query and file interaction.
* Enterprise-Grade Encryption: Data encrypted at rest and in transit.

## 5.3 Collaboration & Integration

* Multi-User Team Workspaces: Enable team collaboration on AI projects.
* Integration with Enterprise Tools: Connect with ERP, CRM, SharePoint, Slack, and other internal apps.
* Custom APIs & SDKs: Build custom AI workflows and connectors.

## 5.4 Deployment & Scalability

* Runs on GPU Infrastructure: Compatible with dedicated GPUs, GPU clouds you control, or on-prem servers.
* Edge Deployment: Bring AI to remote or offline environments.
* Cluster Scaling: Support for distributed deployments for large organizations.

## 5.5 Future AI Features

* fKnowledge Graph Integration: Build structured insights from unstructured data.
* Predictive Analytics & Business Insights: AI-driven recommendations and forecasting.