Federated AI-Assisted Helpdesk

for UK DRI HPC Facilities

Improving Efficiency, Governance and Technical Support Across a Federated Compute Landscape

Dr. Basden | Dr. Fawada Qaiser

Durham University

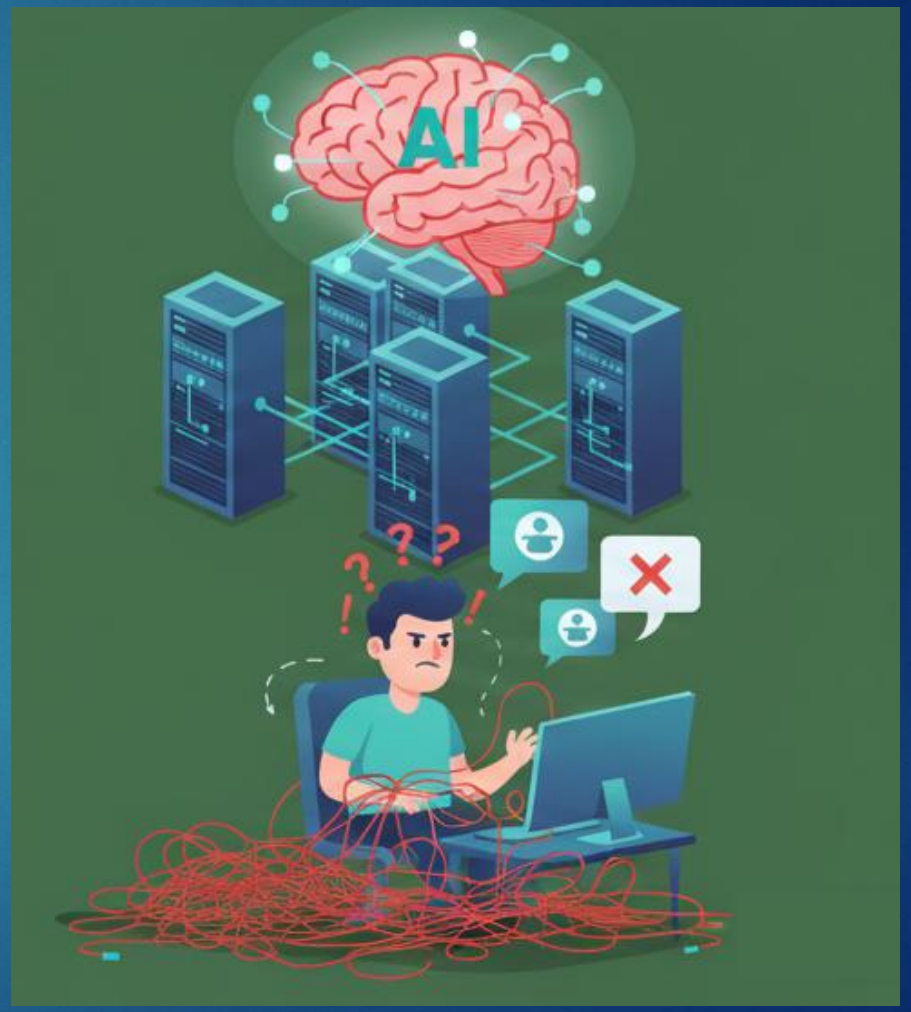
Keele University

National Federated Compute Services NetworkPlus

The Context

HPC and AI: A Paradox

- HPC underpins AI and Large Language Models
- Increasing system complexity across DRI facilities
- Growing user demand and support requests
- Yet internal support processes remain largely manual



The Problem

Current Challenges in HPC Support

- First-line helpdesk queries consume significant time
- No single operator can master all subsystems
- Documentation varies in format and quality
- Knowledge is often siloed within individuals

Proposed Solution

A Federated AI-Assisted Helpdesk

- Automatic first-level support across DRI facilities
- Site-specific, context-aware responses
- Federated architecture respecting site sovereignty
- Cost-effective and scalable

Project Objectives

What We Aim to Demonstrate

- Practical feasibility of AI-assisted helpdesk
- RAG-based site-specific AI responses
- Secure handling of sensitive documentation
- Ethical and governance framework
- Prototype federated web-based front-end

Overall Approach

High-Level Workflow

- Convert documentation into AI-friendly format
- Encode documentation using Retrieval Augmented Generation (RAG)
- Develop site-specific AI assistants
- Address governance, security and ethics
- Deliver federated demonstrator

Work Package Overview

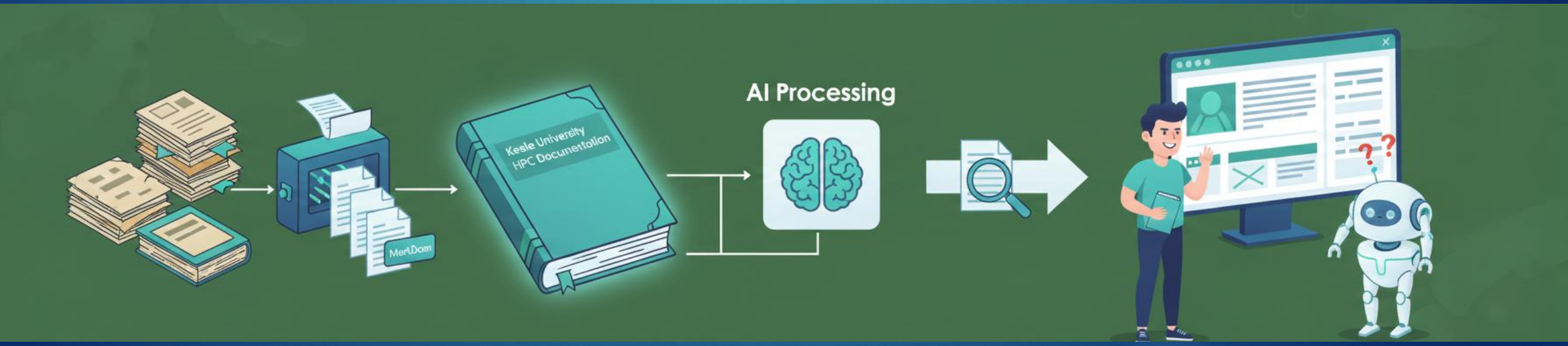
Project Structure

- WP1: AI-Friendly Documentation
- WP2: RAG Feasibility Demonstration
- WP3: Internal RTP Assistant
- WP4: Landscape Survey & Governance
- WP5: Federated Web Front-End

WP1: AI-Friendly Documentation

Preparing Documentation for AI

- Develop guidance for AI-ready documentation
- Convert traditional documentation to structured formats (e.g., Markdown)
- Case study using Keele University HPC documentation
- Improve clarity for both AI and human users



WP2: RAG Feasibility Demonstration

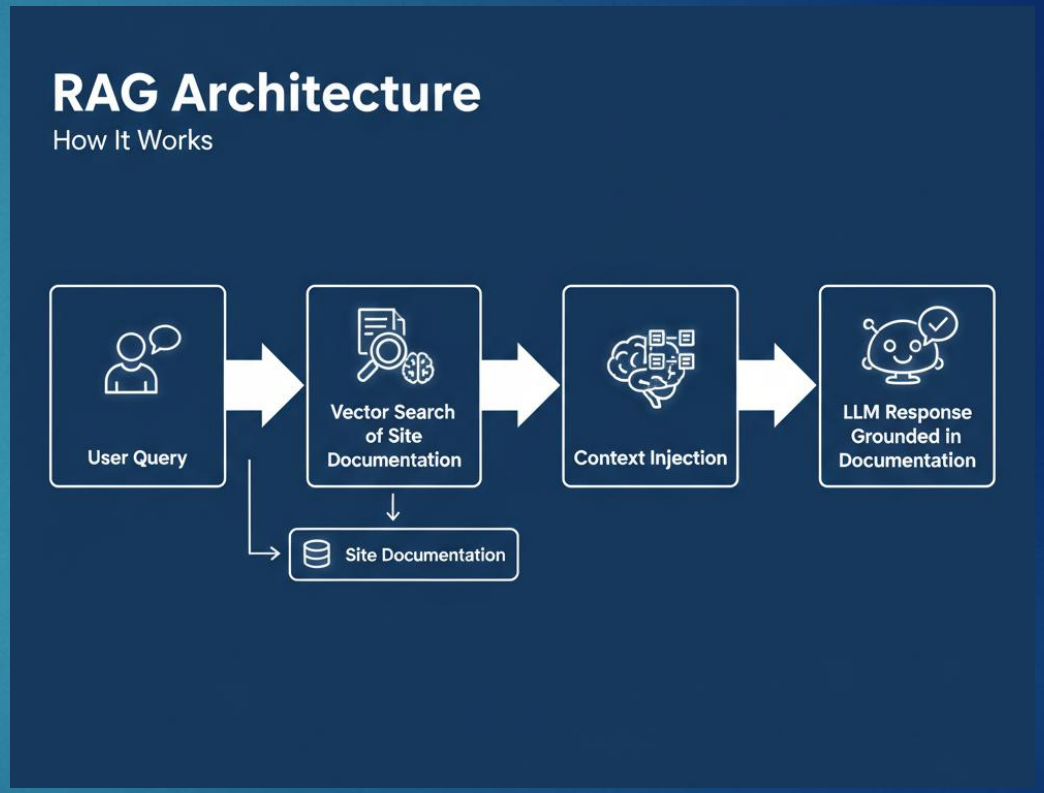
Site-Specific AI Helpdesk Prototype

- Encode multiple HPC documentation sets
- Use Retrieval Augmented Generation
- On-premise deployment using open-source LLMs
- Avoid ongoing subscription costs

RAG Architecture

How It Works

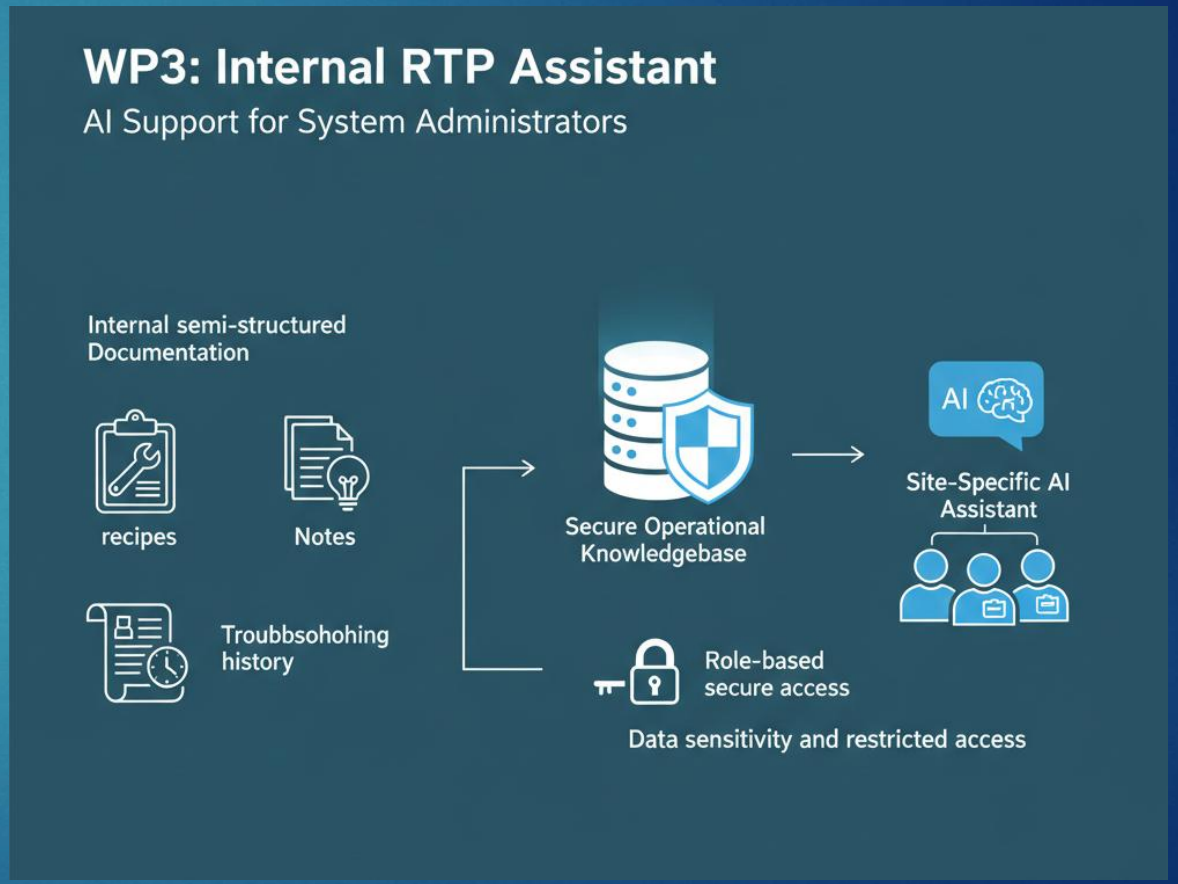
- User Query →
- Vector Search of Site Documentation →
- Context Injection →
- LLM Response Grounded in Documentation



WP3: Internal RTP Assistant

AI Support for System Administrators

- Based on internal, semi-structured documentation
- Includes recipes, notes, troubleshooting history
- Handles sensitive operational knowledge
- Role-based secure access



Security & Information Sovereignty

Protecting Sensitive Information

- On-premise model hosting
- No external data transmission
- Role-based access controls
- Permission-aware query handling

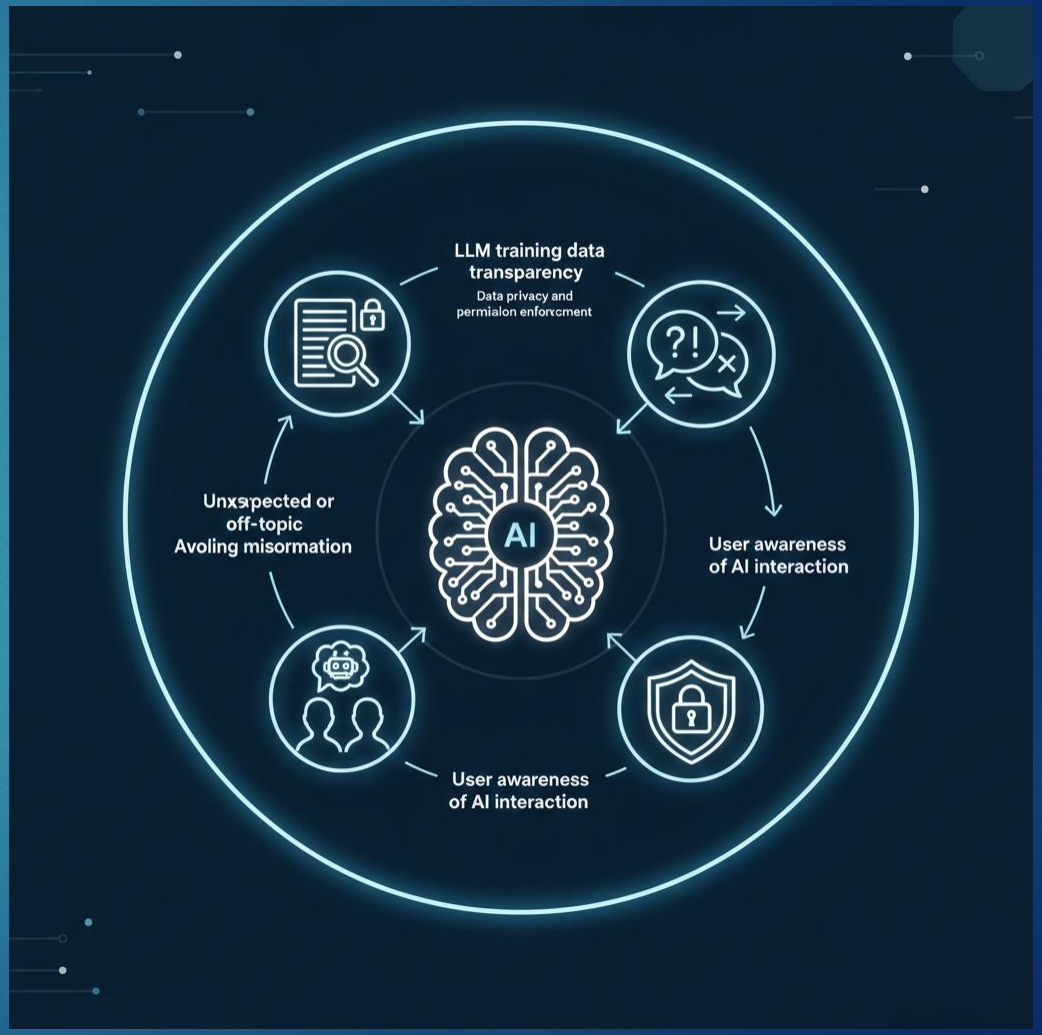
WP4: Landscape Survey

Understanding the Current State

- Survey DRI helpdesk operations
- Identify AI adoption levels
- Document common challenges
- Inform federated roadmap

Ethical & Governance Issues

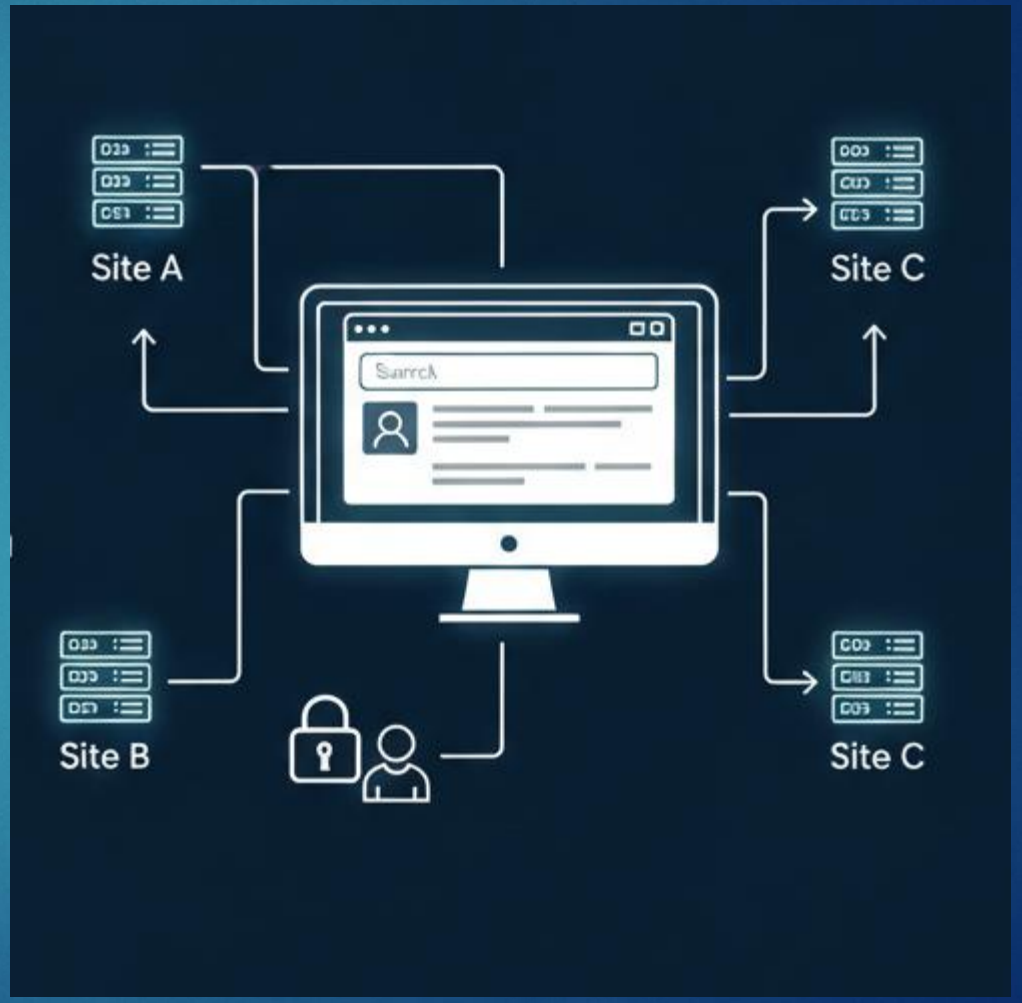
- **Key Considerations**
- LLM training data transparency
- Unexpected or off-topic outputs
- User awareness of AI interaction
- Data privacy and permission enforcement
- Avoiding misinformation or offence



WP5: Federated Web-Based Front-End

Proof-of-Concept Demonstrator

- Web-based federated interface
- Site-specific query routing
- Conversation history
- Role-based protected access
- Feedback loop into documentation



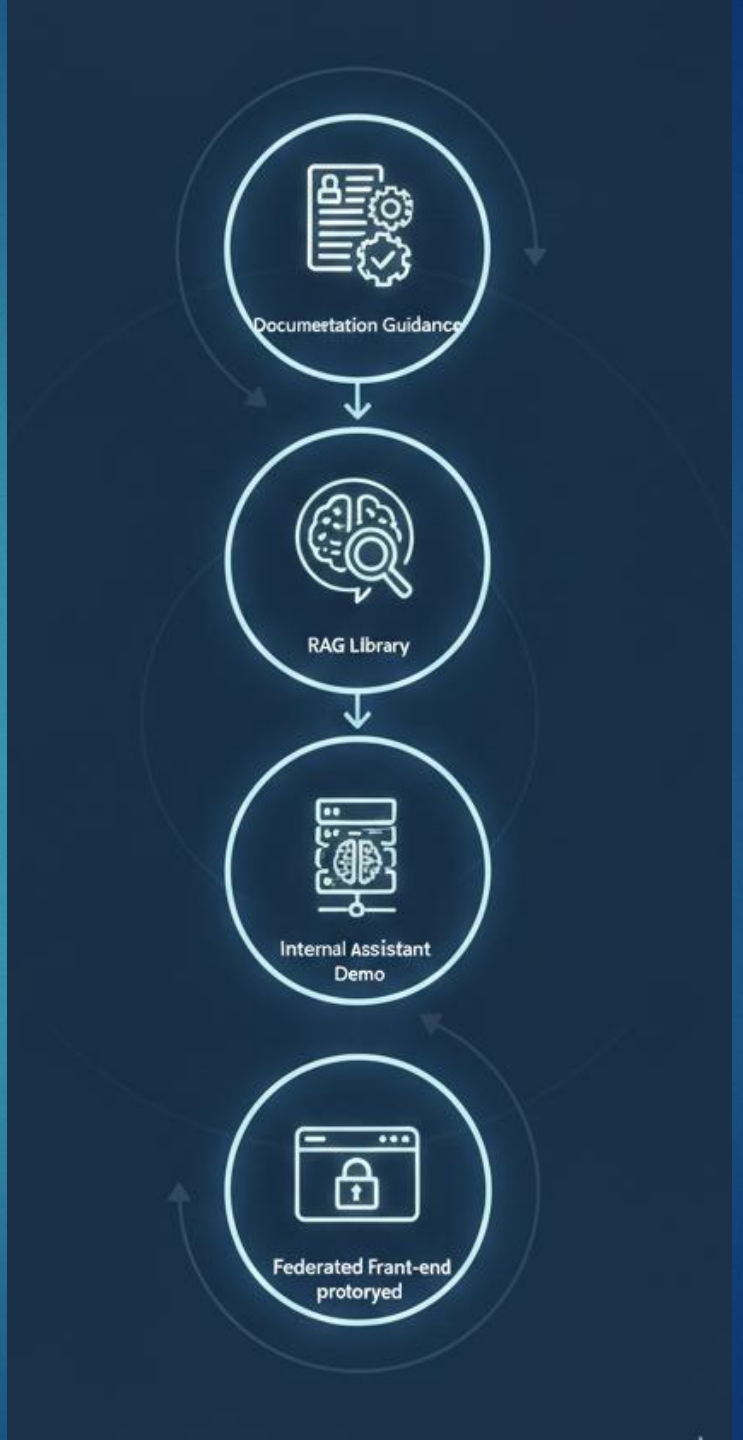
Feedback Loop Model

Continuous Improvement

- User Queries →
- AI Responses →
- Identify Documentation Gaps →
- Improve Documentation →
- Better Future Responses

Key Milestones

- AI-friendly documentation guidance completed
- RAG vector library operational
- Internal assistant demonstration
- Governance survey completed
- Federated front-end prototype delivered



Expected Outcomes

- Federated AI-assisted helpdesk prototype
- AI-friendly documentation framework
- Governance and ethical guidelines
- Productivity and efficiency gains
- Roadmap input for UK DRI federation

Thank you