Database Management & Integration

ChatMED Project

How do we build databases that feed our LLMs the massive data they need — while ensuring absolute security and compliance?

Database Architecture Fundamentals

Goal

High availability, data integrity, and strict confidentiality

Challenge

Integrating unstructured medical data — text, histories, imaging — for LLMs

Key Use Case

Secure, rapid data availability for Retrieval-Augmented Generation (RAG)

Security Posture & Controlled Access

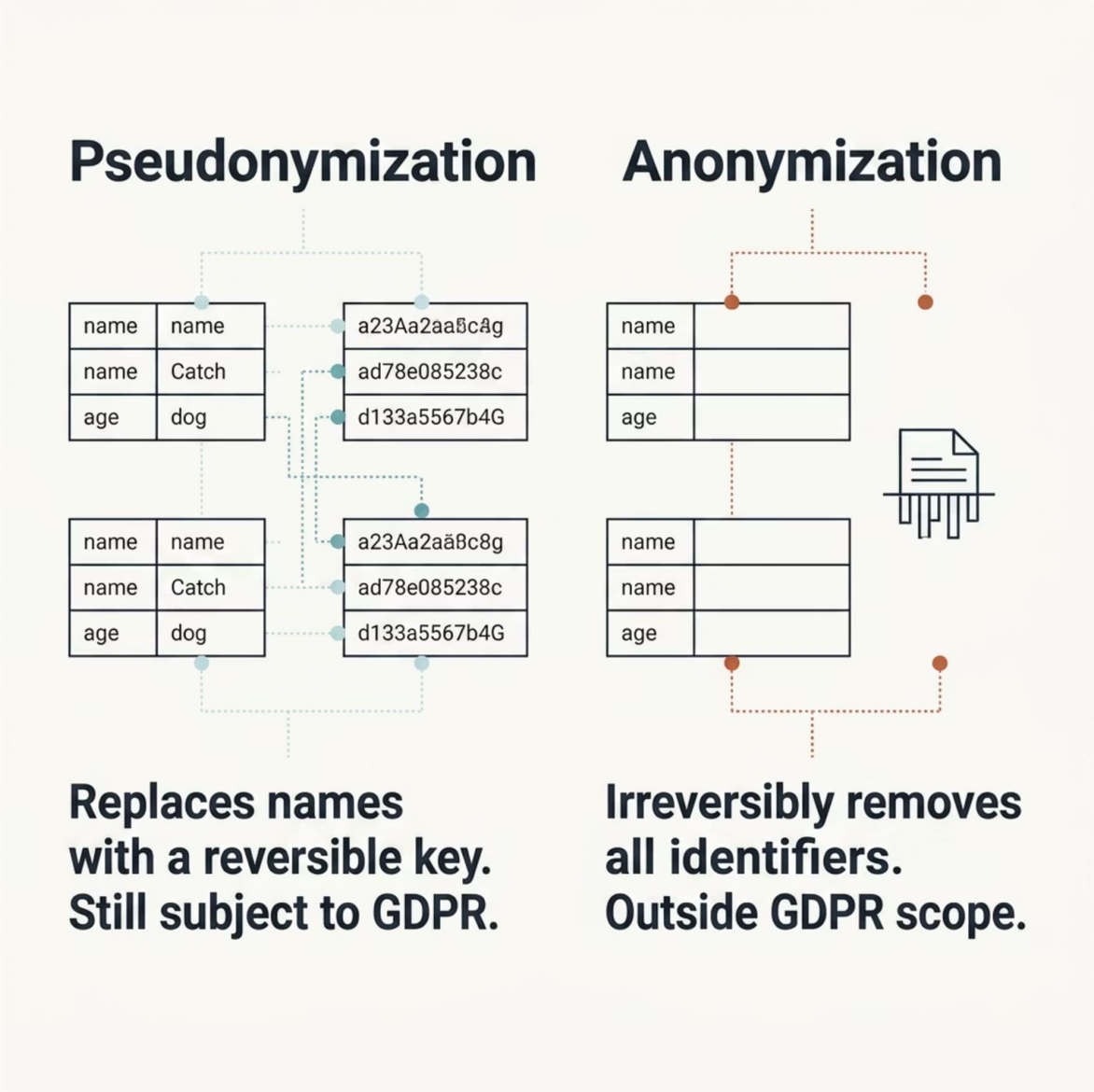
Role-Based Access Control (RBAC)

Researchers and clinicians hold different database permissions — access is defined by role, not convenience.

Principle of Least Privilege

The AI model itself only queries the data required for the specific prompt — nothing more.

Anonymization vs. Pseudonymization



This is the most critical technical distinction. Pseudonymization merely masks data — it remains regulated under GDPR.

True anonymization **irreversibly breaks the link** to the patient. For AI training without regulatory burden, data must be truly anonymized.

The Medical Data Paradox

1

AI Training

Demands massive, detailed datasets

2

Privacy Law

Demands data minimization

3

Solution

Strategic Data Segregation — physically and logically separating operational from training data

HOMEDOCTOR

HomeDOCtor: Database Architecture

RAG Backend

A secure, static vector database of verified Slovenian medical guidelines powers all responses.

Ephemeral User Cache

User prompts are held in temporary cache, processed, then **deleted**. No persistent user database means a drastically reduced attack surface.

HOMEDOCTOR

HomeDOCtor: Ensuring Quality Data

Verified Data Source

Only verified medical protocols are integrated — preventing the LLM from hallucinating harmful advice.

Continuous Updates

The database is continuously updated to reflect current OTC drug availability and clinical guidelines.

For HomeDOCtor to be safe, the underlying database must be flawless.

NEUROLOGY ORCHESTRATOR

Neurology Orchestrator: Database Needs

Multi-Modal Integration

Requires integration with hospital EHRs and multi-modal data: MRI, CT, and EEG readings.

Persistent Encrypted Storage

Unlike HomeDOCtor, the Orchestrator demands robust, persistent, and encrypted storage solutions.

NEUROLOGY ORCHESTRATOR

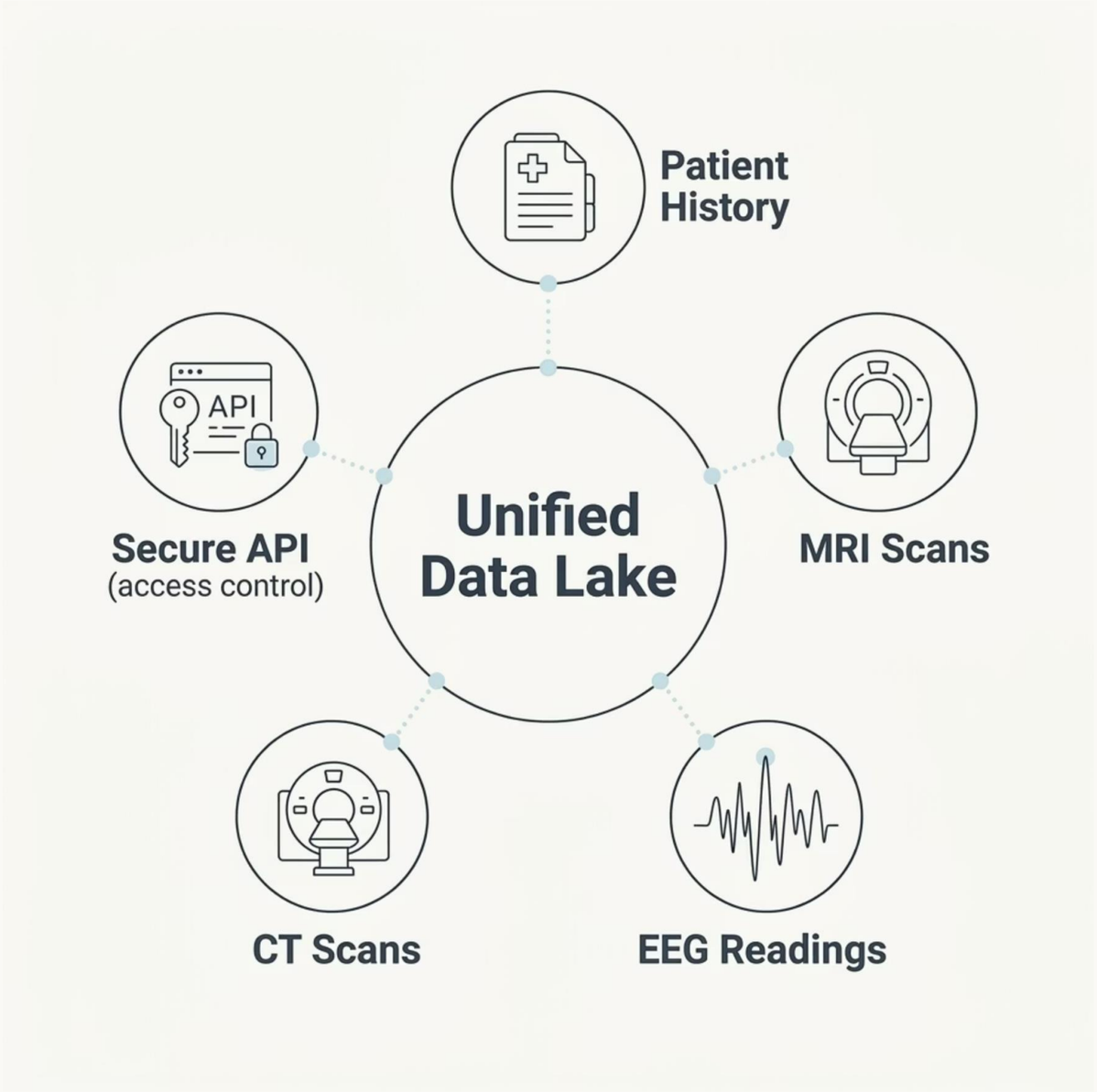
Multi-Modal Integration Challenges

The Challenge

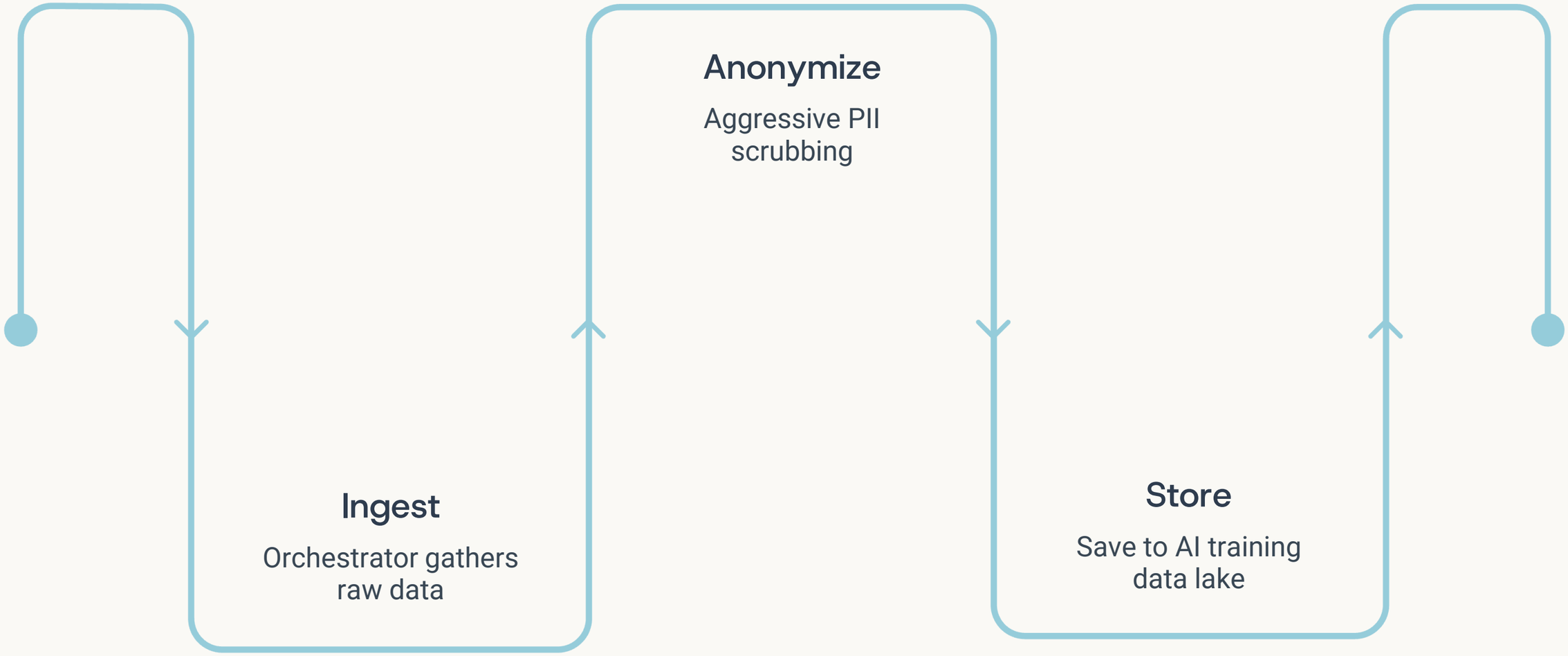
Combining unstructured text (patient history) with visual data (MRI scans) in a single coherent pipeline.

The Solution

Unified data lakes with access tightly controlled via **secure APIs** — ensuring the AI receives both modalities seamlessly.



Collecting Data for Future Models



To build better models tomorrow, we must build the pipeline today — ingesting, scrubbing, and safely storing data in a separate training lake.

Anonymizing Clinical Text

Named Entity Recognition (NER)

AI automatically detects and masks names, dates, and locations before text reaches the training database.

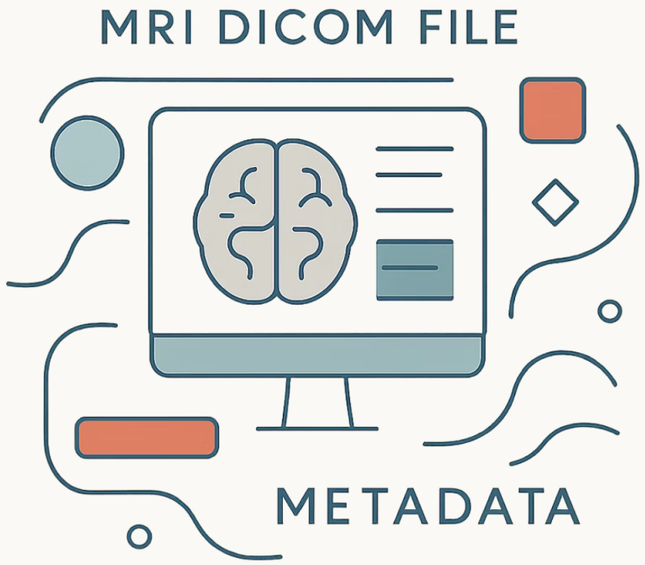
Data Masking

Sensitive fields are replaced with tokens or redacted values throughout clinical notes.

Synthetic Data Generation

Realistic but entirely fabricated patient records supplement training sets without privacy risk.

Anonymizing Medical Images



The Challenge

DICOM files contain hidden metadata headers with patient identifiers — and physical markers on scans can also reveal identity.

The Solution

Header scrubbing strips metadata automatically upon ingestion. **Defacing algorithms** remove physical identifiers from neurological scans.

Preparing for EHDS

European Health Data Space (EHDS) will require health data interoperability across the EU.

Requirement

Cross-border health data interoperability

Standard

Implementing **FHIR** (Fast Healthcare Interoperability Resources)

Outcome

ChatMED tools ready to integrate with hospitals across Europe

Auditing & Logging

Immutable Access Logs

Every database access — by a human or AI agent — is recorded and tamper-proof.

Compliance & Traceability

If an AI makes a bad recommendation, we must be able to audit the **exact data it read**. Logs prove compliance and trace AI decision pathways.

Conclusion: Security by Design

Anonymize Early & Aggressively

Break the link to the patient before data enters any training pipeline.

Segregate Training from Operational Data

Physical and logical separation is non-negotiable.

Build for Interoperability

FHIR-ready databases ensure ChatMED scales across Europe.

Secure database management is the foundation of ethical AI — build it right, and clinicians will trust it.