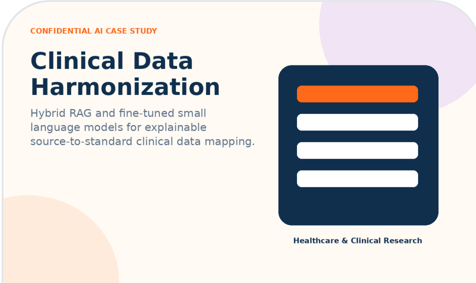
CONFIDENTIAL AI CASE STUDY

Clinical Data Harmonization

Hybrid RAG and fine-tuned small language models for explainable source-to-standard clinical data mapping.

A healthcare / clinical research organization needed to harmonize complex source datasets into a standardized data model for downstream analytics, reporting, and research workflows. The source data contained a large number of variables with inconsistent naming conventions, incomplete descriptions, mixed data types, missing value codes, and domain-specific clinical meanings.

> **Project classification:** This proof of concept was developed for a confidential healthcare and clinical research organisation.



BYOND BOUNDRYS

HEALTHCARE & CLINICAL RESEARCH

INDUSTRY	CLIENT TYPE	PROJECT STAGE
Healthcare & Clinical Research	Confidential healthcare / clinical research organisation	Proof of concept
EVIDENCE	**DELIVERY SCOPE**	
Pilot-observed	Hybrid retrieval, fine-tuning and validation	

Reduced source-to-standard mapping preparation effort

The prototype reduced the preparation required before experts reviewed source-to-standard mappings.

Generated structured and reviewable harmonisation plans

It produced structured harmonisation plans that domain experts could inspect, correct and approve.

Improved candidate retrieval through hybrid RAG

Hybrid semantic, keyword and metadata retrieval improved candidate selection when variable names did not match directly.

01 / BUSINESS CONTEXT

Who needed the solution

A healthcare / clinical research organization needed to harmonize complex source datasets into a standardized data model for downstream analytics, reporting, and research workflows. The source data contained a large number of variables with inconsistent naming conventions, incomplete descriptions, mixed data types, missing value codes, and domain-specific clinical meanings. Manual harmonization required significant domain expertise and was time-consuming, especially when variables required mapping, renaming, transformation logic, category conversion, or new derived variable creation. To protect confidentiality, client name, dataset name, project name, and standard model details are not disclosed.

---

02 / CHALLENGE

What needed to change

The organization wanted to reduce the manual effort involved in mapping raw clinical variables to a target standard. The main challenges were: • Source variables were often ambiguous or poorly documented. • Similar clinical concepts appeared with different names across datasets. • Some variables required direct mapping, while others required transformation logic. • Traditional keyword matching was not enough because clinical meaning depends on context. • LLM outputs needed control because they could over-predict mappings or generate unnecessary transformations. • The solution needed to be generalized, not hardcoded for one dataset.

Ambiguous or incomplete source metadata	Source variables were often ambiguous or poorly documented.
Different names for the same clinical concept	Similar clinical concepts appeared with different names across datasets.
Direct mappings and transformation logic in the same workflow	Some variables required direct mapping, while others required transformation logic.
Keyword-only matching missed clinical context	Traditional keyword matching was not enough because clinical meaning depends on context.

Uncontrolled model outputs could over-predict operations	LLM outputs needed control because they could over-predict mappings or generate unnecessary transformations.

03 / WORKFLOW TRANSFORMATION

Before and after

Previous workflow	Structured workflow
• Ambiguous or incomplete source metadata • Different names for the same clinical concept • Direct mappings and transformation logic in the same workflow • Keyword-only matching missed clinical context • Uncontrolled model outputs could over-predict operations	• Enrich source metadata • Retrieve hybrid candidates • Generate mapping plans • Validate and score outputs • Review structured outputs

04 / SOLUTION

How we approached it

We designed an AI-assisted harmonization workflow using hybrid RAG, metadata enrichment, and fine-tuned small language models. Instead of relying on a single large model, the solution used multiple specialized small LLM components fine-tuned for different parts of the workflow.

01

Enrich source metadata

Profile names, descriptions, types and value patterns before retrieval.

02

Retrieve hybrid candidates

Combine semantic, keyword, metadata and domain-aware matching.

03

Generate mapping plans

Use fine-tuned small language models for mappings and transformations.

04

Validate and score outputs

Apply schema checks, guardrails, confidence filters and repeatable evaluation.

---

05 / EXAMPLE WORKFLOW

From input to reviewable output

1 **Enrich source metadata** →

2 **Retrieve hybrid candidates** →

3 **Generate mapping plans** →

4 **Validate and score outputs**

5 **Run validation and evaluation** →

6 **Domain-expert review**

---

06 / DELIVERY SCOPE

What the delivery covered

Workflow and domain mapping

- Metadata profiling and enrichment
- Candidate target-variable retrieval
- Structured harmonisation plans
- Expert review workflow

AI, retrieval and evaluation

- Hybrid semantic and BM25 retrieval
- Fine-tuned small language models
- Schema validation and confidence filtering
- Batch evaluation and benchmarking

Data and engineering

- Python processing pipeline
- FAISS vector retrieval
- PyTorch and LoRA fine-tuning
- Pandas and NumPy data handling

Scope boundary

The work was limited to a proof of concept focused on retrieval, mapping and validation. Production integration, the final target standard and externally approved benchmark reporting were outside this public scope.

How the system is organised

- Python
- FAISS
- BM25
- Pydantic
- PyTorch
- Pandas
- LoRA
- NumPy

Schema and operation validation

Structured schemas, consistency checks and transformation rules reduce malformed or unnecessary operations.

Confidence-aware review

Low-confidence mappings remain visible for domain-expert review rather than being silently accepted.

Source traceability

Candidate mappings retain the metadata and retrieved target context used to form the recommendation.

Confidentiality boundary

Client, dataset and target-model details remain anonymised in the public case study.

Value observed during validation

1 **Reduced source-to-standard mapping preparation effort** — Pilot-observed

The prototype reduced the preparation required before experts reviewed source-to-standard mappings.

2 **Generated structured and reviewable harmonisation plans** — Pilot-observed

It produced structured harmonisation plans that domain experts could inspect, correct and approve.

3 **Improved candidate retrieval through hybrid RAG** — Pilot-observed

Hybrid semantic, keyword and metadata retrieval improved candidate selection when variable names did not match directly.

4 **Reduced unsupported operations through validation and guardrails** — Pilot-observed

Schema checks and guardrails reduced unsupported mappings and unnecessary transformation steps.

5 **Created a repeatable benchmark and batch-evaluation workflow** — Pilot-observed

The workflow created a repeatable benchmark and batch-evaluation process for further model iteration.

Technology and component roles

PY **Python**
AI, data-processing and backend logic

FA **FAISS**
Fast vector-similarity retrieval

BM **BM25**
Keyword and exact-term retrieval

PY **Pydantic**
Structured schema and output validation

PY **PyTorch**
Model training and inference

PA **Pandas**
Tabular data processing and analysis

LO **LoRA**
Parameter-efficient model fine-tuning

NU **NumPy**
Numerical processing

---

10 / RELATED WORK

Explore related case studies

INSURANCE & COMPLIANCE

InsureAI

RAG-based policy review, real-time knowledge retrieval and automated policy-gap analysis.

FINANCIAL SERVICES & ANALYTICS

FinSight AI

Multi-tenant bank-statement analysis, financial Q&A and visual intelligence for finance teams.

CYBERSECURITY & COMPLIANCE

CMMC Compliance AI Backend

Agentic RAG, NIST-grounded reasoning and structured compliance-document generation for CMMC workflows.

Working on a similar challenge?

Discuss the workflow, data, integrations and evidence required for a practical first release.