OncoLlama:

Using Large Language Models (LLMs) to Quantify Cancer Inequalities at Population Scale

Building scalable, high-quality cancer data curation to transform and improve our understanding and provision of cancer care

National Network Driver Project (London)

ONE LONDON | LONDON SECURE DATA ENVIRONMENT

Part of the NHS Research Secure Data Environment Network

Project Leads

Dr. Joe Zhang (AI Centre for Value Based Healthcare, GSTT), Dr. Emily Jin (AI Centre for Value Based Healthcare, GSTT), Dr. Lawrence Adams (AI Centre for Value Based Healthcare, GSTT), Dr. Martin Chapman (AI Centre for Value Based Healthcare, GSTT)

Partners

NHS Guy's and St Thomas' NHS Foundation Trust · AI Centre for Value Based Healthcare · King's College London · NHS The Leeds Teaching Hospitals NHS Trust · NHS Cambridge University Hospitals NHS Foundation Trust · NHS University College London Hospitals NHS Foundation Trust · NHS East Kent Hospitals University NHS Foundation Trust · aws · Meta · flatiron. · ZS · IgniteData

Introduction

Each year, over 3 million people with cancer in England receive care through the NHS, which results in a large amount of clinical notes. These documents (e.g., clinician correspondence and multidisciplinary team meeting summaries to diagnostic and test reports) contain detailed insights into patients' diagnoses, genetic biomarkers, disease progression, treatments, and their care outcomes. However, as these usually take the form of unstructured text records written in everyday clinical language, they are missed in coded datasets typically used for research and service planning. This gap in care records severely limits the ability to identify cancer patterns, improve services, and match patients to clinical trials.

Timeline and Impact

OncoLlama is built to be portable and widely shareable. The aim is for any NHS cancer centre to adopt it easily through a tailored deployment toolkit, clear documentation, and comprehensive training resources. This will demonstrate how large-scale collaboration can be achieved across the NHS SDE Network.

Three studies are also being carried out:

- **Study One**: OncoLlama-generated data will be validated and used in a joint analysis of cancer inequalities across populations to understand differences in diagnosis, access to treatment including trials, and outcomes
- **Study Two**: demonstrating how OncoLlama technology can improve the completeness of national cancer records
- **Study Three**: exemplifying how enriched data can accelerate and broaden patient identification and workflows in clinical trials, potentially opening up access to innovative treatments for marginalised populations

Project Summary and Outputs

OncoLlama is an NHS-developed AI tool that reads clinical documents and automatically extracts detailed cancer information with over 98.5% accuracy across 20 types of cancer in adults. Rather than relying on expensive and time-consuming manual curation, OncoLlama can be run inside NHS firewalls to process millions of cancer records into comprehensive cancer datasets safely and efficiently.

Operates Safely within NHS Trust Infrastructure

Enables the rapid creation of registry-quality Real-World Datasets and patient identification for clinical trials, while maintaining data privacy

Extracts Detailed Cancer Information

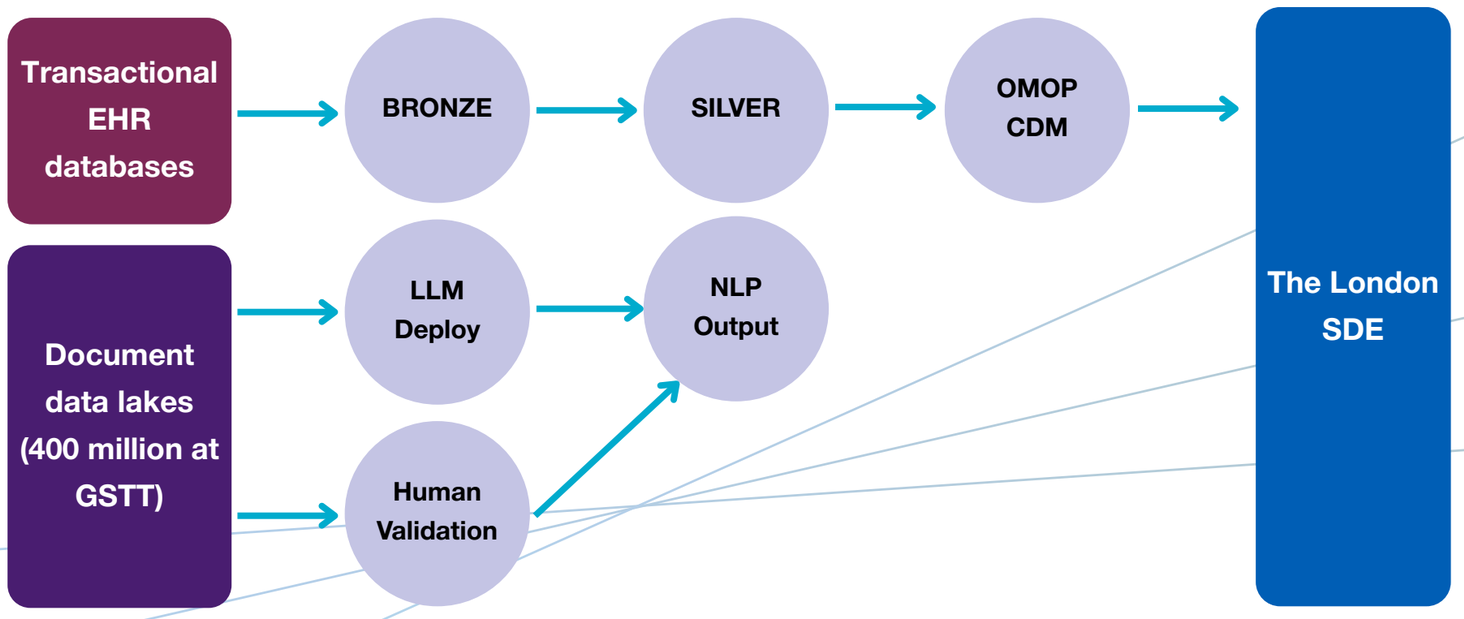
Including diagnosis date, topography, morphology, biomarkers, staging, metastases, treatment/toxicity/progression events etc.

Quick, Proven and Efficient Processing

Currently deployed into production, with over half a million documents processed at Guys' and St Thomas' NHS Foundation Trust (GSTT)

Building on OncoLlama's successful rollout at GSTT, this project brings together five leading cancer centres from London, Cambridge, Leeds, and Kent-Medway to explore how real-world data can support research and speed up clinical trials.

Pipeline Architecture in Deployment (GSTT)*



**Pictured is the basic pipeline architecture for OncoLlama in deployment at Guy's and St. Thomas' NHS Foundation Trust (GSTT), linking with structured data in OMOP and feeding into the London SDE (5.4 million patients).*