

# From radiology reports to early prognostic markers: benchmarking LLMs in chronic liver disease

Hania Paverd<sup>1,2,3,4</sup>, Zeyu Gao<sup>1,4</sup>, Golnar Mahani<sup>1,4</sup>, Margarete A. Fabre<sup>2,3,5</sup>, Sarah Burge<sup>4</sup>, Matthew Hoare<sup>1,3,4\*</sup>, Mireia Crispin-Ortuzar<sup>1,4\*</sup>

<sup>1</sup>Early Cancer Institute, University of Cambridge, Cambridge, UK <sup>2</sup>Centre for Genomics Research, Discovery Sciences, BioPharmaceuticals R&D, AstraZeneca, Cambridge, UK <sup>3</sup>Cambridge University Hospitals NHS Foundation Trust, Cambridge, UK <sup>4</sup>Cancer Research UK Cambridge Centre, University of Cambridge, Cambridge, UK <sup>5</sup>Department of Haematology, Wellcome-MRC Cambridge Stem Cell Institute, Jeffrey Cheah Biomedical Centre, University of Cambridge, Cambridge, UK \* Co-senior authors

**Key message: Open-source LLMs with constrained decoding enable robust, scalable, automated data extraction from complex, multi-source clinical records.**

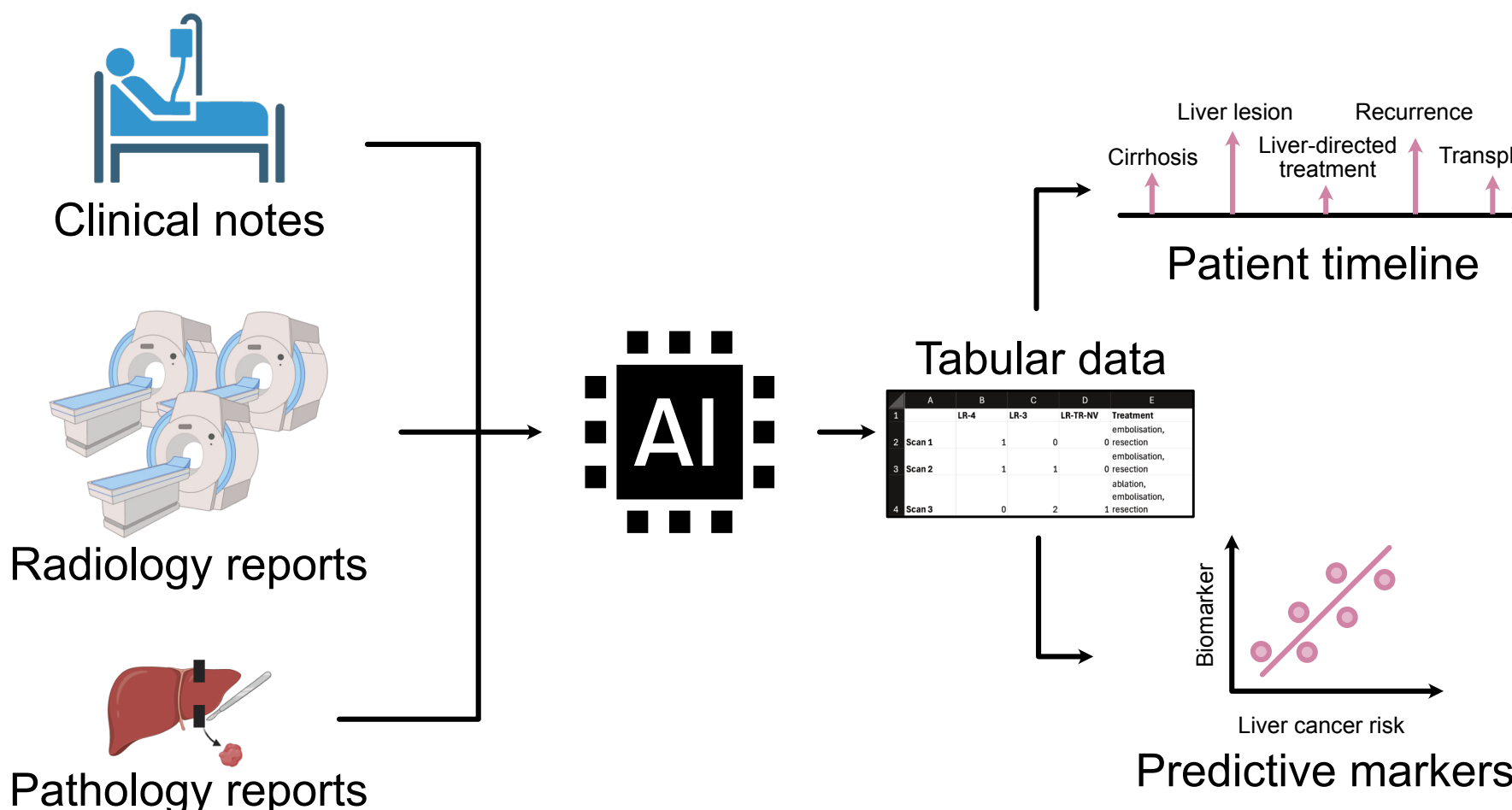


**Contact:** [hmv56@cam.ac.uk](mailto:hmv56@cam.ac.uk)



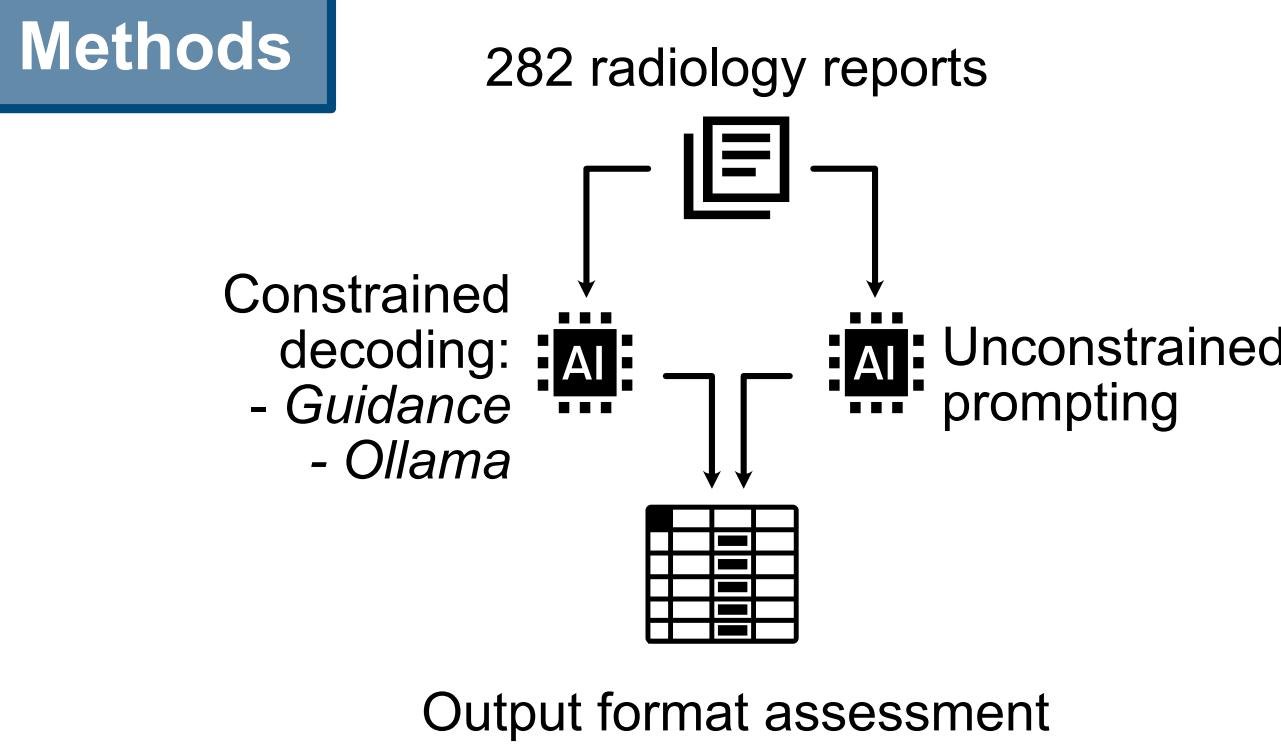
## Vision & Goals

- Electronic health records contain rich clinical data but are often stored as **unstructured text**, making manual curation labor-intensive and error-prone.
- Large language models (LLMs) can **automate extraction** of structured data from clinical reports, yet most studies focus on a **single modality**, use **proprietary models**, and overlook challenges like **output formatting**.



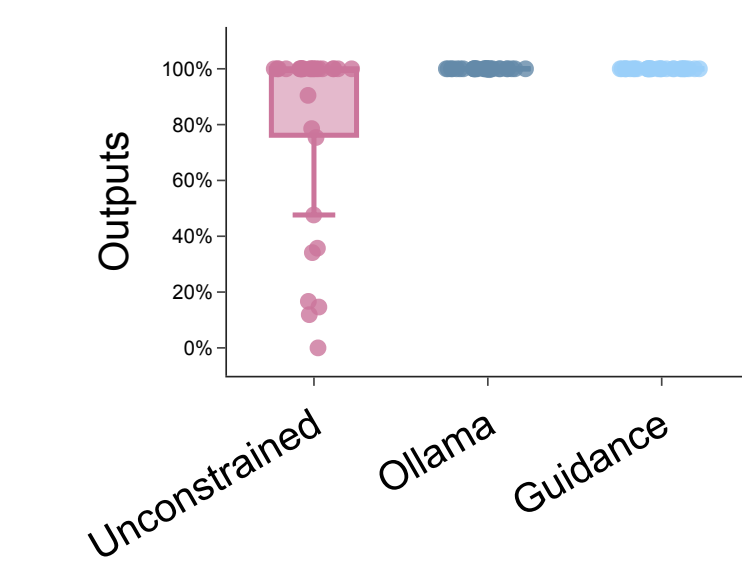
- This study benchmarks **open-source LLMs** and **constrained decoding** methods across **multiple report types** to reconstruct **disease trajectories** in liver transplant patients.

## Does constrained decoding help?



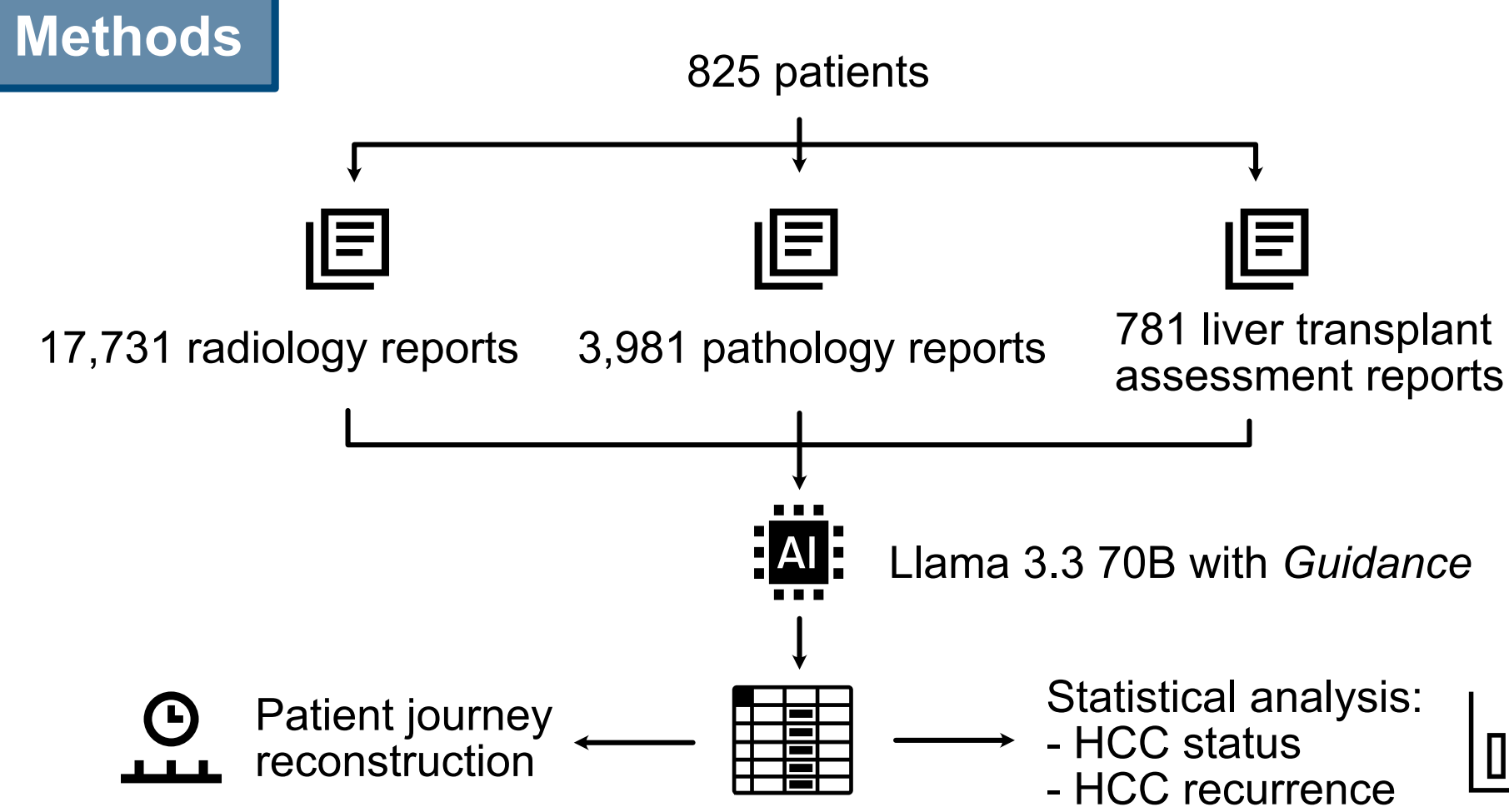
## Results

### Valid outputs per question



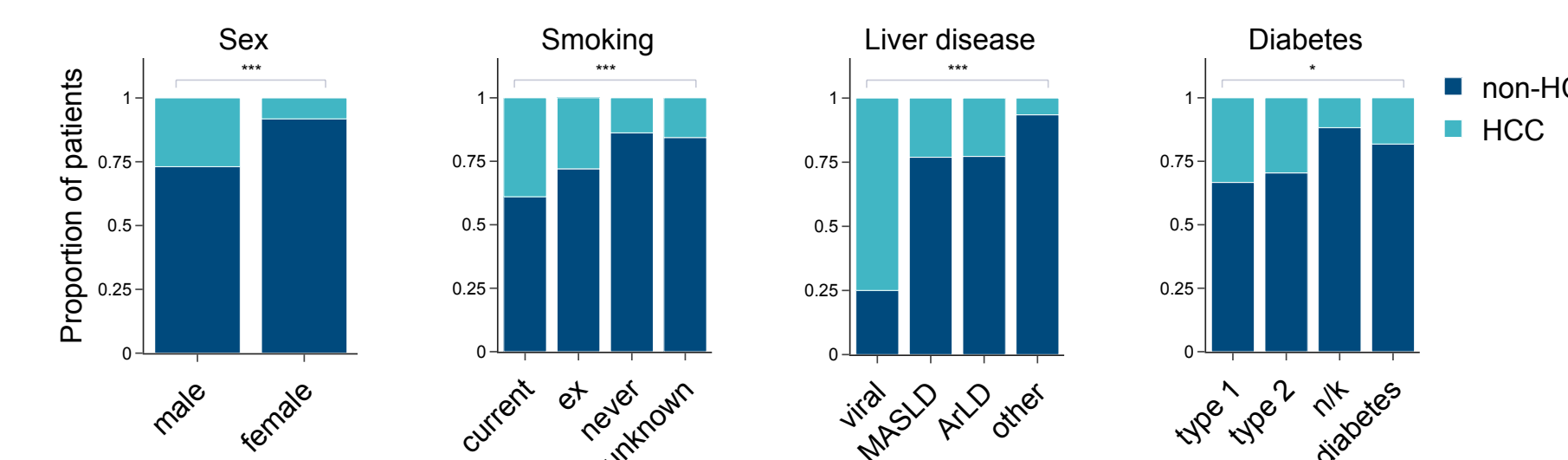
**Key finding: Constrained decoding leads to perfect output format adherence.**

## What can we learn from LLM-extracted data?



## Results

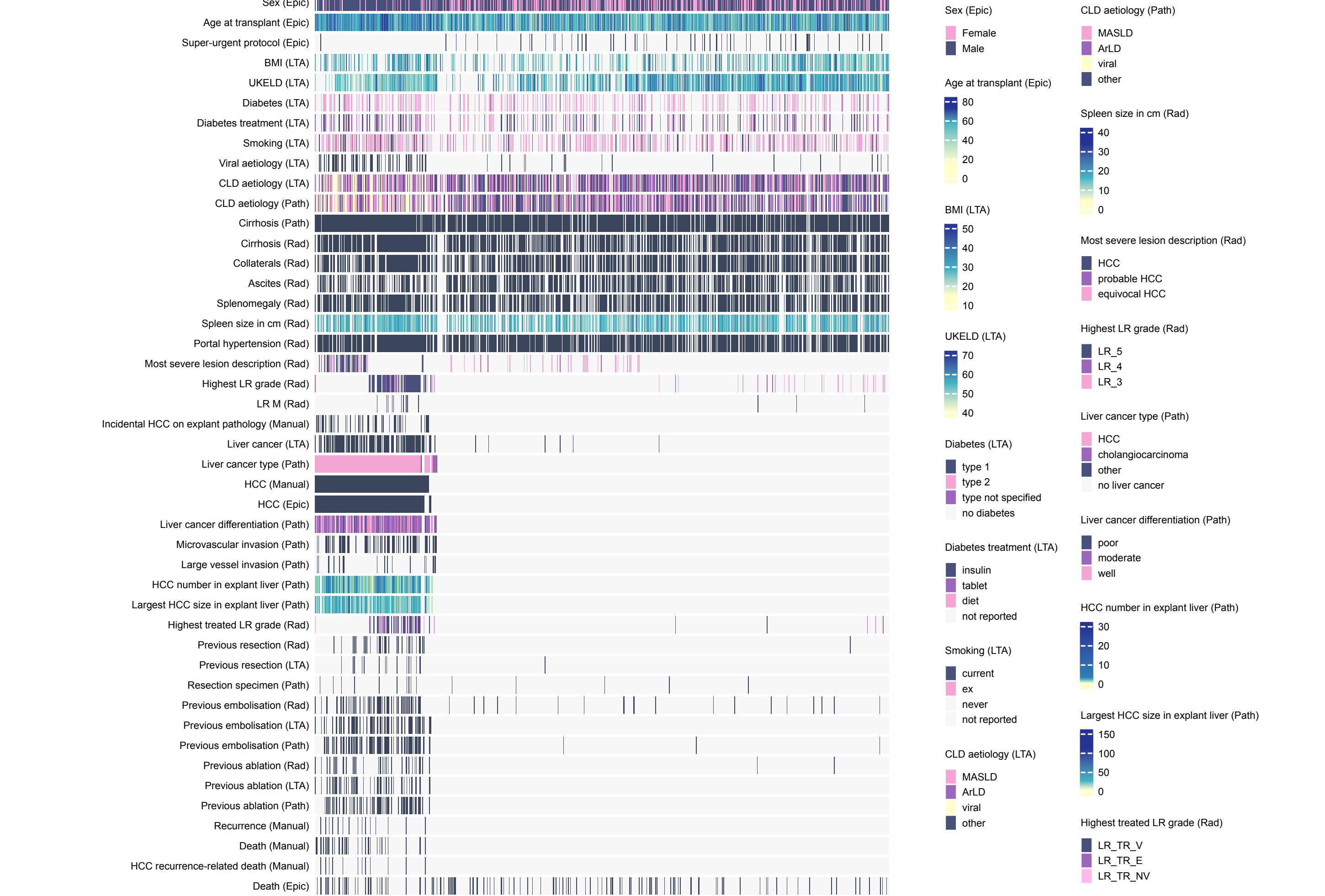
### Predictive markers



**Key finding: LLMs can identify clinically-validated risk factors of liver cancer.**

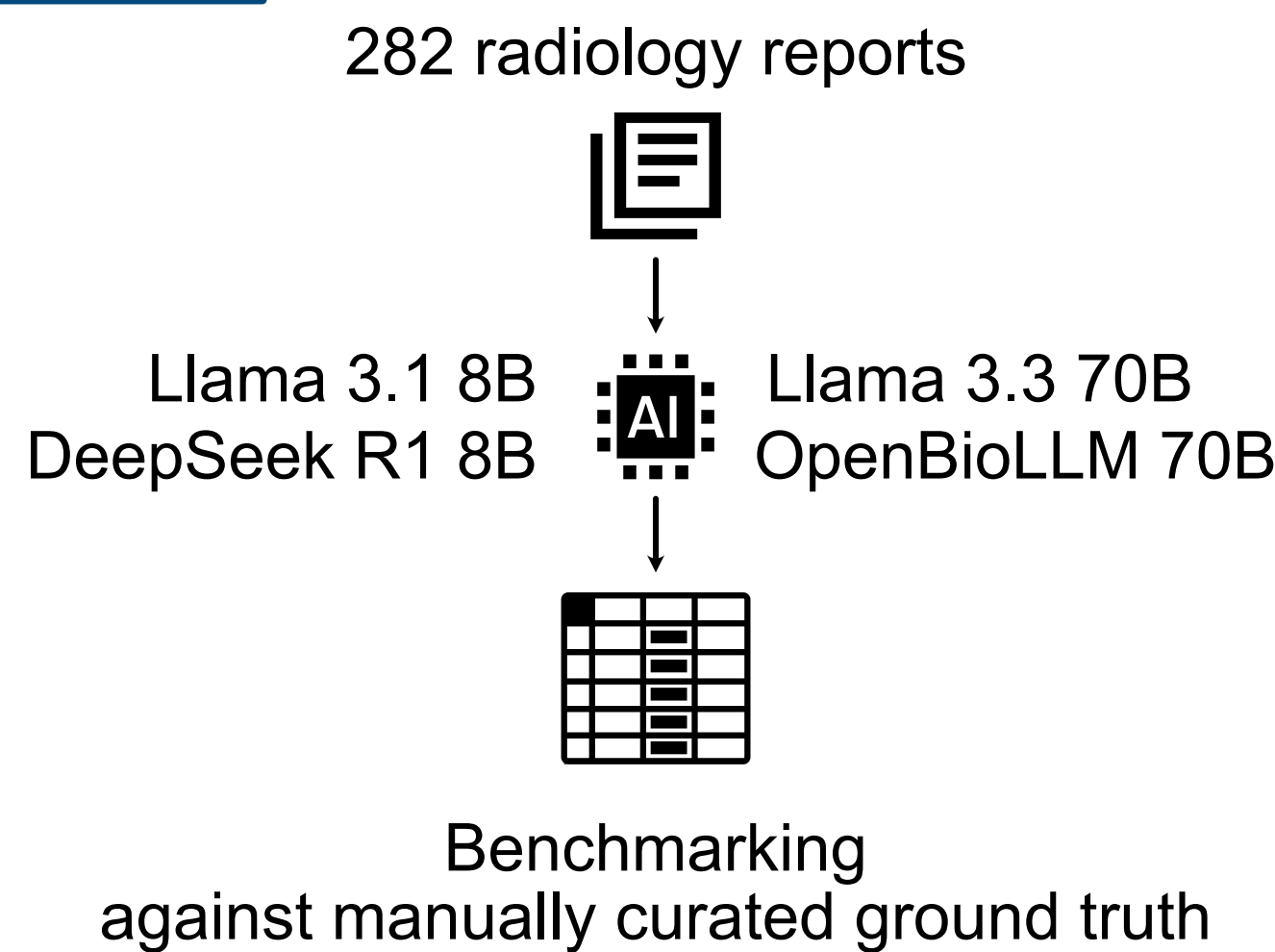
## Results

### Summative database



## How do LLMs perform?

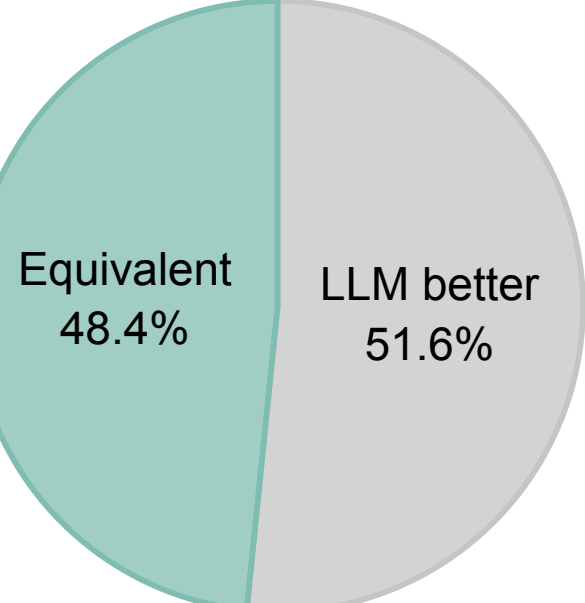
## Methods



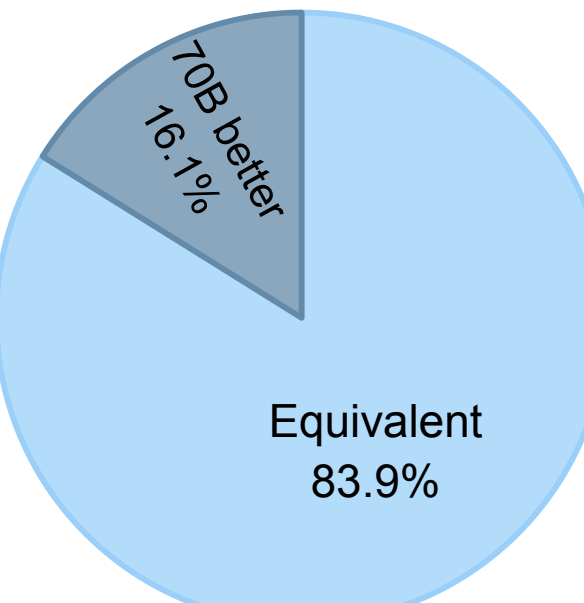
**Key finding: Llama 70B outperforms regex, OpenBioLLM and Llama 8B.**

## Results

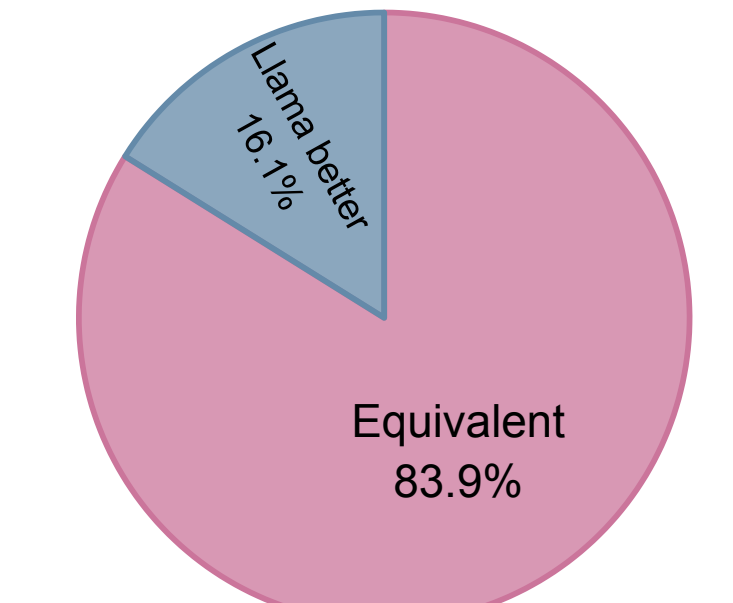
### Regex vs LLM



### Llama 70B vs 8B

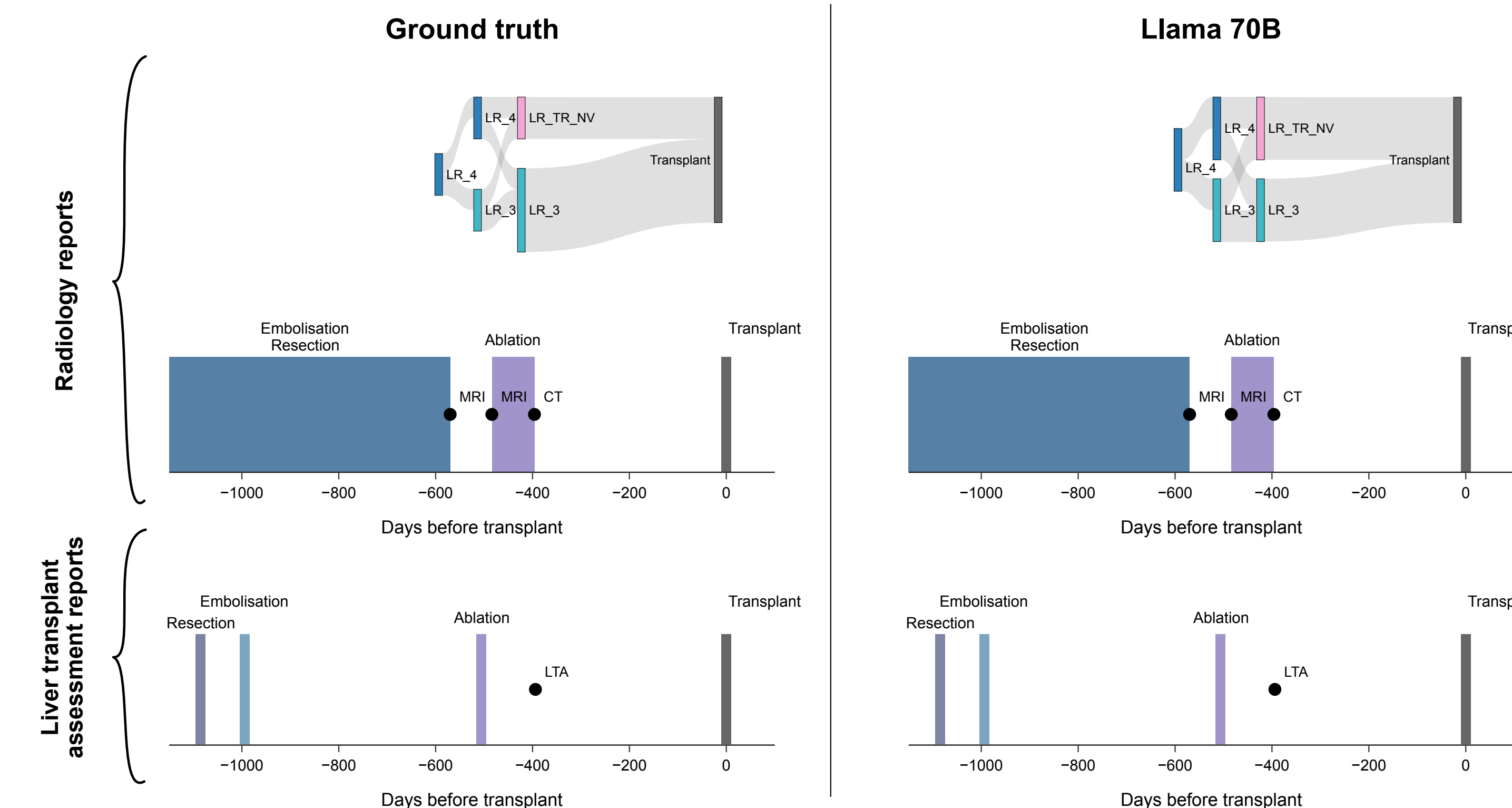


### OpenBioLLM vs Llama



## Results

### Timeline reconstruction



**Key finding: LLM-extracted data can serve to reconstruct patient timelines from multiple report modalities.**

## Results

### Timeline evaluation

