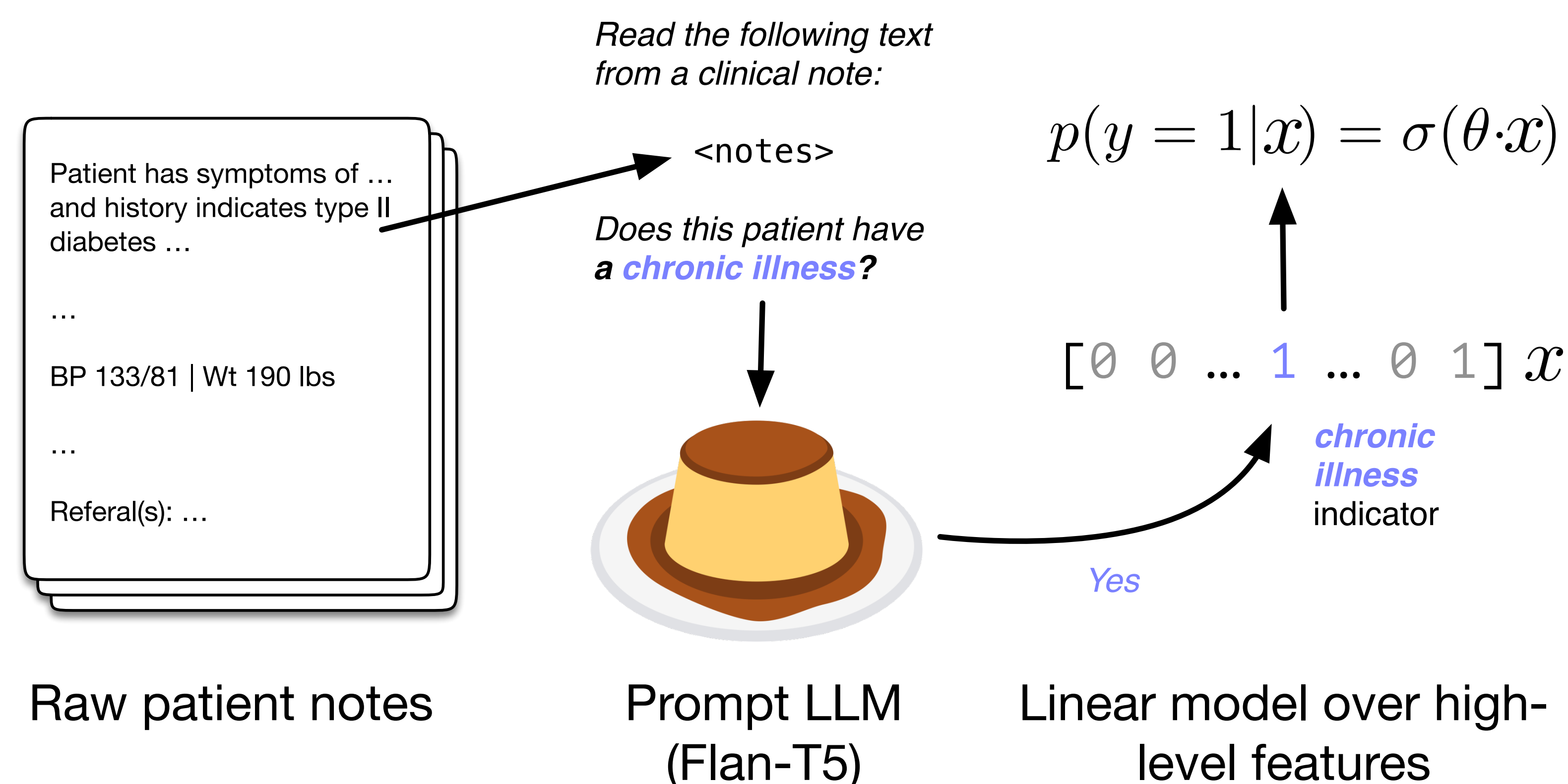# CHiLL: Zero-shot Custom Interpretable Feature Extraction from Clinical Notes with Large Language Models

Denis Jered McInerney[1], Geoffrey Young[2], Jan-Willem van de Meent[3], Byron C. Wallace[1]

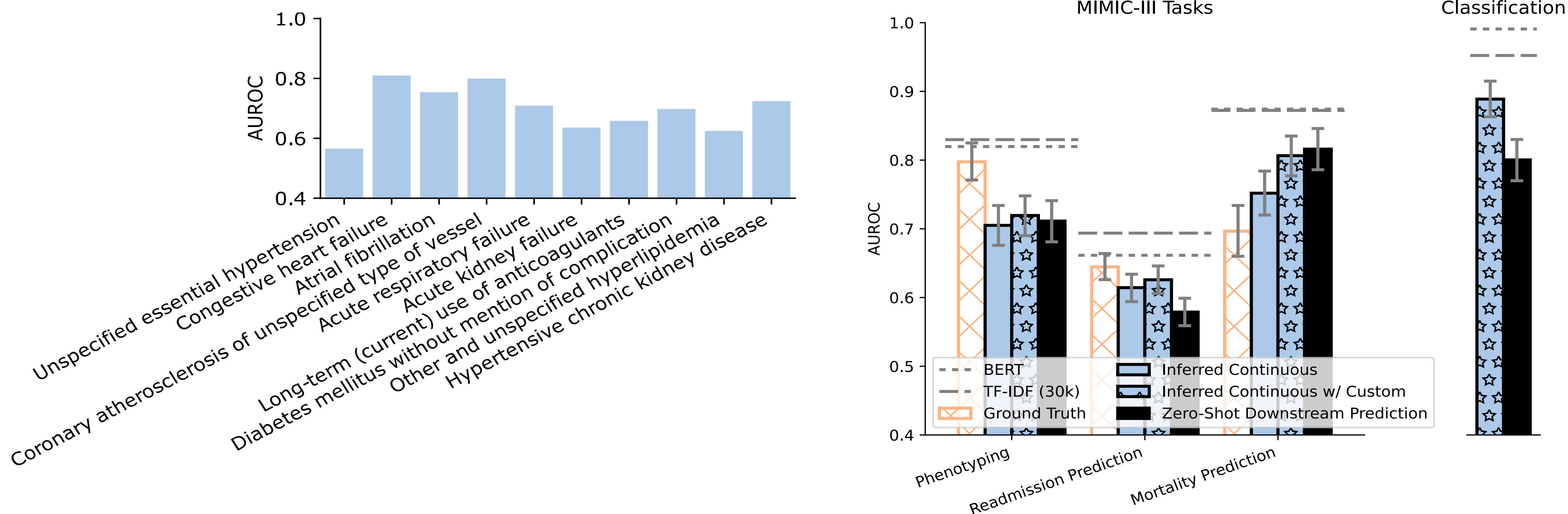[1]Northeastern University, [2]Brigham and Women's Hospital, [3]University of Amsterdam

Introduction

Can we **improve interpretability** by automating **high-level feature extraction** with Large Language Models for use in linear models for downstream tasks?

Read the following text from a clinical note:

<notes>

Does this patient have **a chronic illness**?

$$p(y = 1 | x) = \sigma(\theta \cdot x)$$

$$[0\ 0\ \ldots\ 1\ \ldots\ 0\ 1]\ x$$

chronic illness indicator

Yes

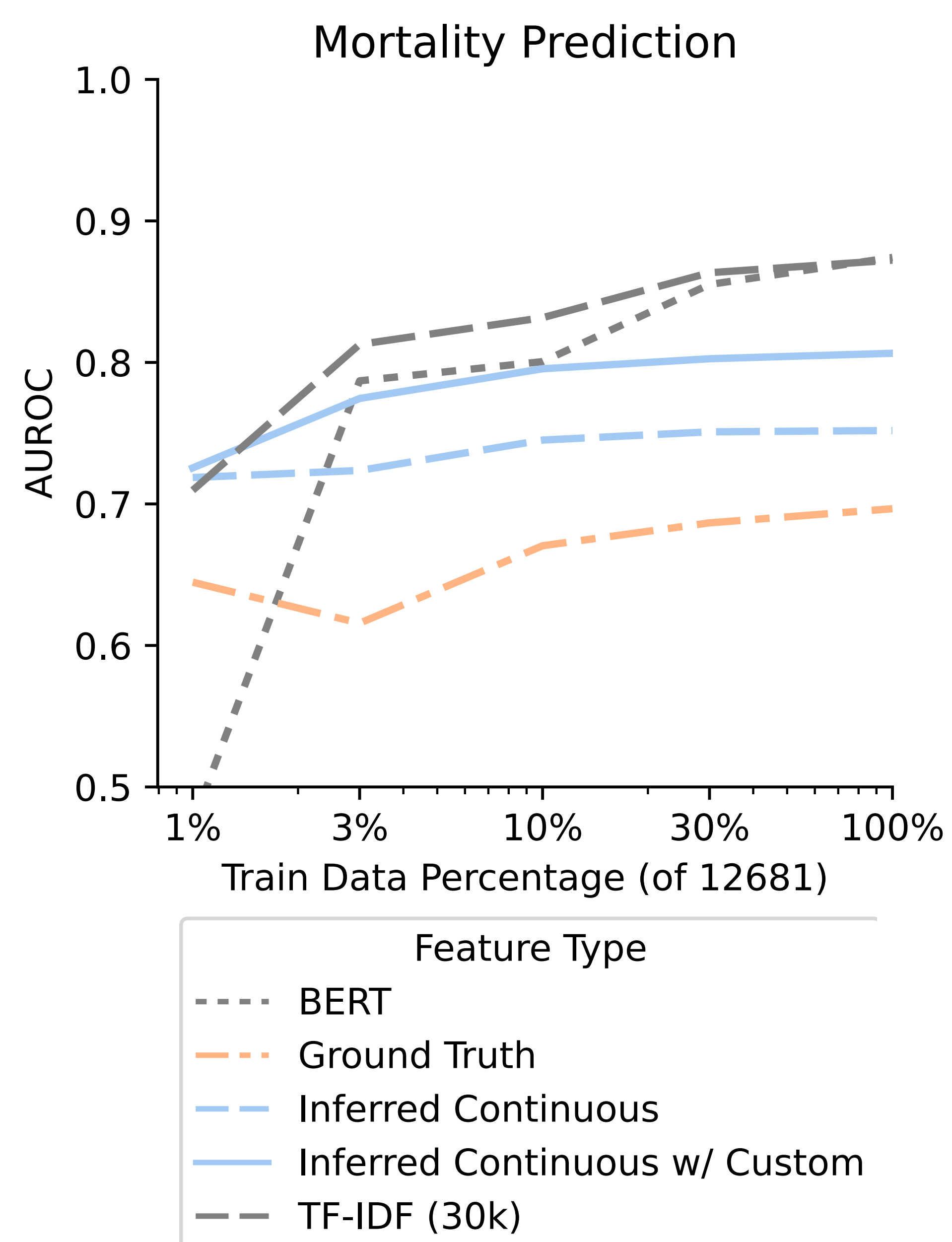Raw patient notes → Prompt LLM (Flan-T5) → Linear model over high-level features

Feature Extraction and Downstream Performance

We find we are able to get reasonable AUC and F1 scores for feature extraction (left), leading to reasonable downstream performance (right). Though this is not up to par with TF-IDF or BERT models, it is competitive with using corresponding ground-truth features.
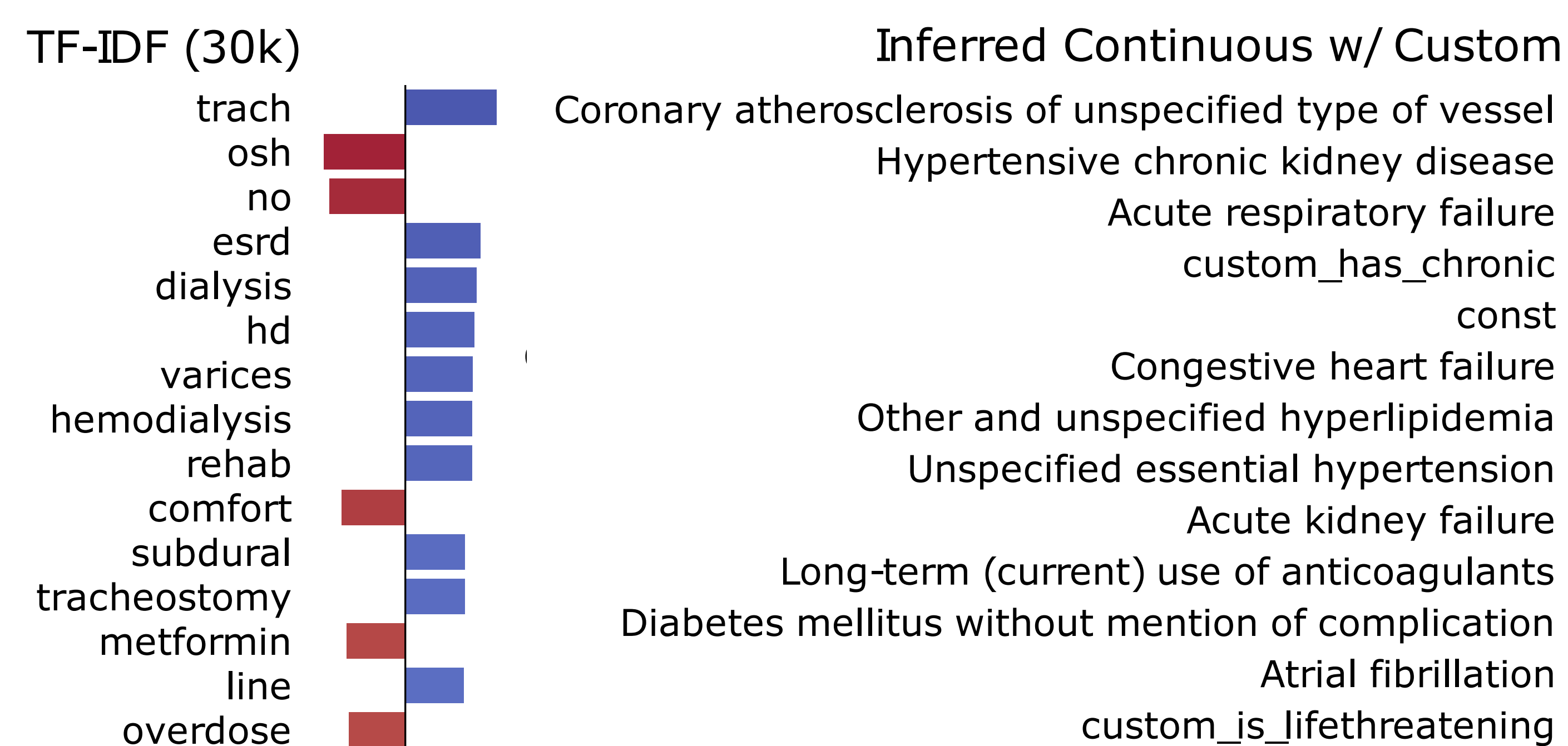

Data Efficiency

We find that our method is much less data-hungry than BERT, and is often competitive with TF-IDF when less train data is available.
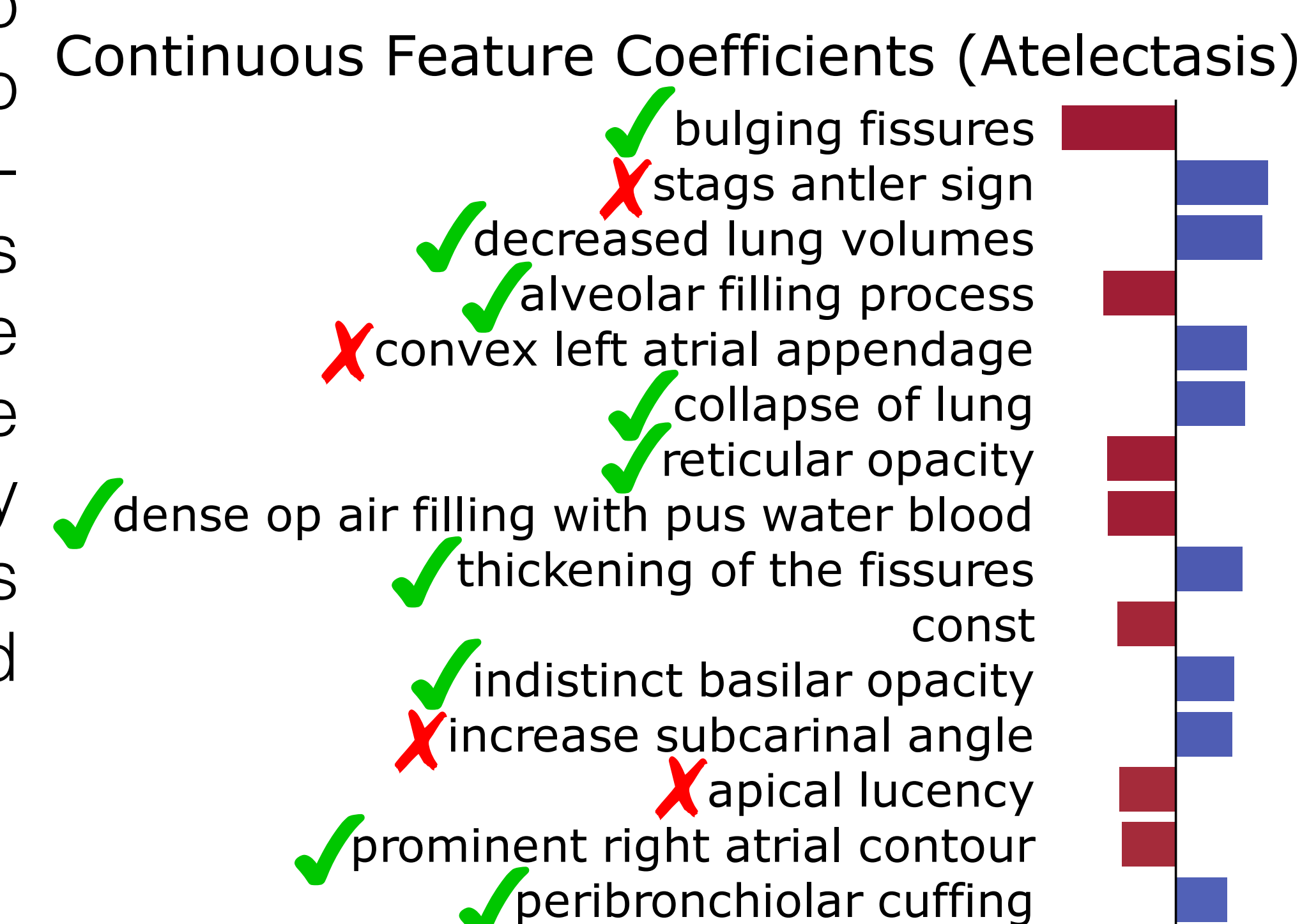
Interpretability

Readmission Prediction Features

Hand-crafted features can also map much more naturally to well-defined concepts than TF-IDF features (left). The weights that models produce are therefore more interpretable and do seem to qualitatively align with clinical judgments according to expert-annotated model weights (right).

Atelectasis Prediction Features

We see that in specific instances, experts can inspect the extracted features and the features that most impacted the model's decision to evaluate what went right or wrong. In this case, the expert discovered the labeler made a mistake, and the model was right!

HISTORY: Intubated, evaluate ET tube.

FINDINGS: The ET tube is 3.5 cm above the carina. The NG tube tip is off the film, at least in the stomach. Right IJ Cordis tip is in the proximal SVC. The heart size is moderately enlarged. There is ill-defined vasculature and alveolar infiltrate, right greater than left. This is markedly increased compared to the film from two hours prior and likely represents fluid overload.

Targets (CheXpert Labeler): Cardiomegaly, Lung Opacity, Support Devices

Model Predictions: Cardiomegaly, Edema, Lung Opacity, Support Devices

Clinical Judgment: Edema — Should be a target!