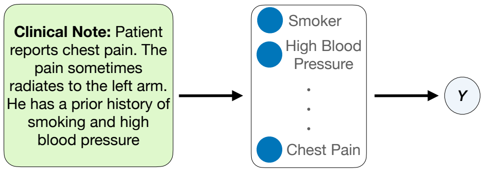
Bayesian Concept Bottleneck Models with LLM Priors

UCSF
ZUCKERBERG SAN FRANCISCO GENERAL Hospital and Trauma Center

Jean Feng Avni Kothari Lucas Zier Chandan Singh Yan Shuo Tan

NUS National University of Singapore
Microsoft

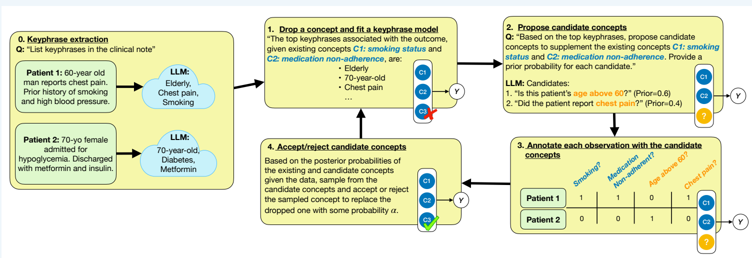


Motivation

- Concept Bottleneck Models (CBMs) leverage black-box models to extract interpretable concepts, which serve as inputs to a transparent prediction model.
- CBMs currently require human experts to identify and extract a set of candidate concepts a priori. The size of this set is limited by practical constraints and, more importantly, may not include truly relevant concepts.

BC-LLM

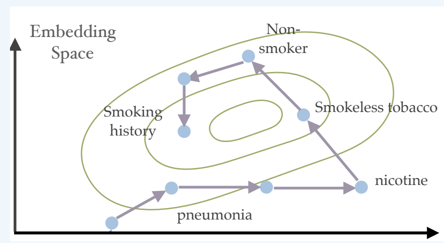
- Leverage LLMs to provide a prior over concepts, propose candidate concepts, extract concepts, and iteratively refine the concepts.
- To overcome errors or inconsistencies in the LLM, BC-LLM frames the LLM-guided concept search as a Bayesian posterior sampling procedure, which allows for statistically rigorous inference and uncertainty quantification:
  - ***Theorem (Informal):*** *Even if the LLM defines an imperfect prior, BC-LLM will converge to the true concepts asymptotically.*



Overview of BC-LLM

Efficient LLM Search over Concepts

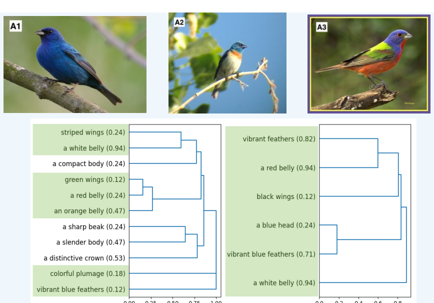
- Formally, BC-LLM searches over candidate concepts using *Gibbs sampling*.
- To search over concepts as efficiently as possible, we leverage *Multiple Try Split-Sample Metropolis-within Gibbs*, *in which* the LLM proposes multiple candidate concepts each iteration and selects the one best aligned with the data.



Example sampling procedure for a single concept

BC-LLM learns accurate yet interpretable concept bottleneck models by using LLMs to iteratively hypothesize, annotate, and refine candidate concepts within a statistically rigorous Bayesian framework

Paper



Left and right dendrograms are concepts learned using 1/3 versus all of the training data. Highlighted concepts are distinguishing bird features.

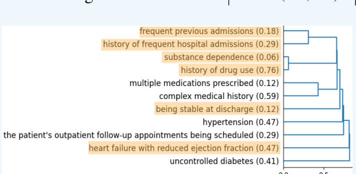
Experiment: Bird type classification

- Task: Learn a CBM to predict bird type. To make the task more difficult and prevent test data leakage, the LLM is not told it is classifying bird types.
- BC-LLM outperforms all other CBM learning procedures and performs as well as, if not better, than black-box models.

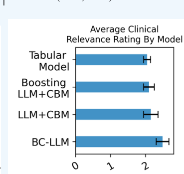
AUCs of CBMs and black box (ResNet) for classifying bird subcategories

	In-distribution			*OOD*
Method	Accuracy (↑)	AUC (↑)	Brier (↓)	Entropy (↑)
BC-LLM	**0.680** (0.614, 0.747)	**0.874** (0.840, 0.907)	**0.428** (0.357, 0.500)	0.865 (0.693, 1.036)
LLM+CBM	0.640 (0.573, 0.707)	0.810 (0.768, 0.853)	0.452 (0.377, 0.528)	0.663 (0.474, 0.852)
Boosting LLM+CBM	0.538 (0.463, 0.614)	0.722 (0.673, 0.772)	0.577 (0.499, 0.654)	0.842 (0.630, 1.054)
Human+CBM	0.658 (0.591, 0.725)	0.835 (0.791, 0.879)	0.499 (0.414, 0.584)	0.758 (0.558, 0.959)
LLM+CBM (No keyphrases)	0.555 (0.488, 0.623)	0.759 (0.713, 0.805)	0.651 (0.548, 0.754)	0.626 (0.495, 0.757)
ResNet	0.664 (0.613, 0.716)	0.853 (0.821, 0.885)	0.457 (0.398, 0.516)	**0.914** (0.748, 1.079)

Method	AUC (95% CI)	Brier (95% CI)
BC-LLM	**0.64** (0.58, 0.70)	**0.14** (0.12, 0.62)
LLM+CBM	0.59 (0.52, 0.65)	0.29 (0.25, 0.33)
Boosting LLM+CBM	0.59 (0.52, 0.66)	**0.14** (0.12, 0.16)
Bag-of-words	0.52 (0.46, 0.58)	0.29 (0.25, 0.34)



Concepts learned by BC-LLM where highlighted concepts received scores from clinicians as being highly predictive



Average clinical ratings for concepts from different methods

Experiment: Revising a Readmission Risk Prediction Model with Clinical Notes

- Task: Determine if there are useful concepts in discharge notes for improving on an existing tabular model for predicting 30-day unplanned readmission risk. Learn a CBM that revises the existing risk prediction by extracting 4 concepts from clinical notes.
- BC-LLM outperforms all other CBM learning procedures.
- Survey results: Clinicians found the BC-LLM model to be
  - More clinically meaningful and interpretable
  - Contained more causally relevant features
  - More actionable, i.e. suggested clinical actions for reducing readmission risk