Transparente KI Methoden in der Medizin

Carsten Eickhoff

EBERHARD KARLS
UNIVERSITÄT
TÜBINGEN

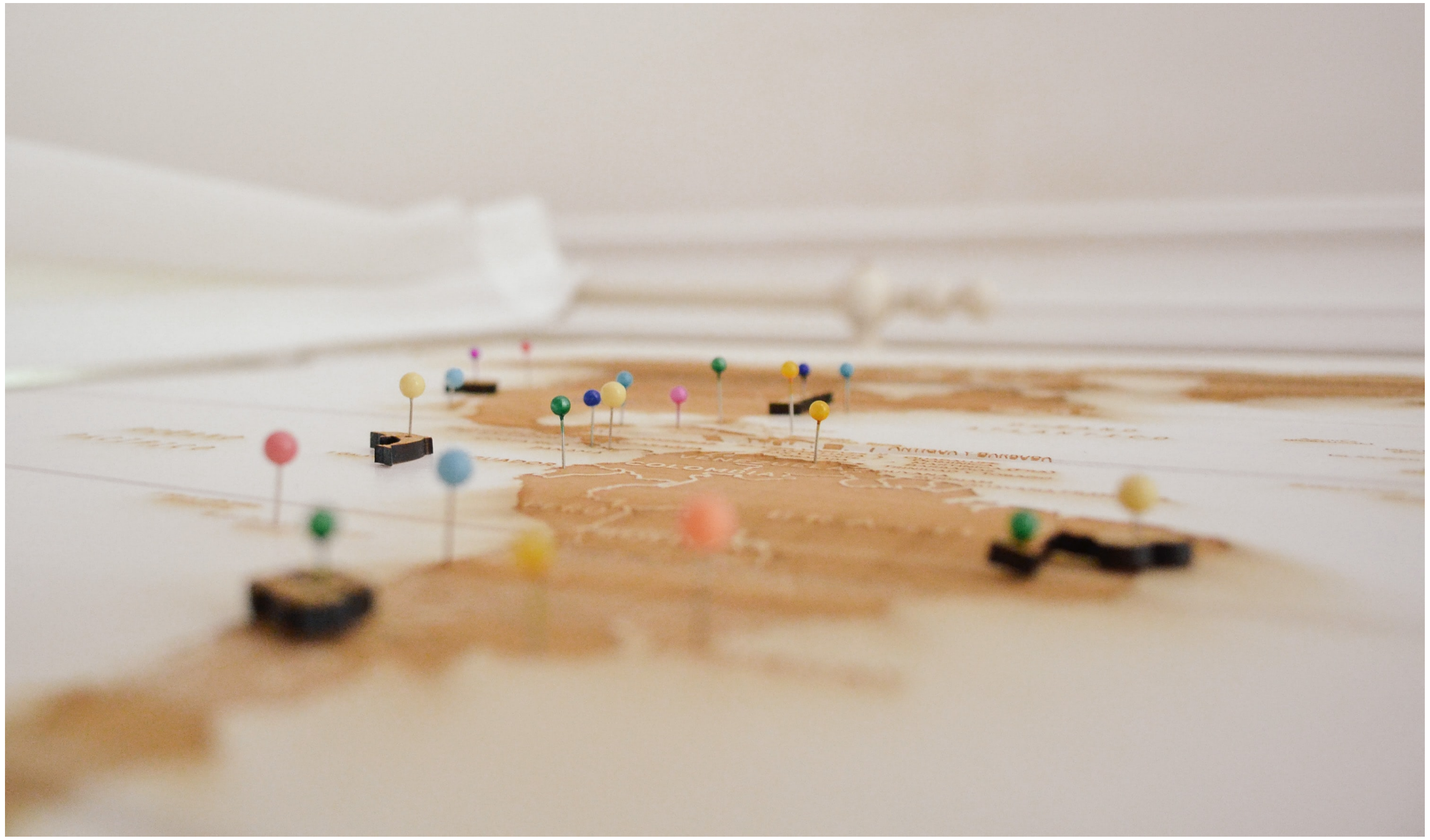
BROWN

Overview

1. **Background, Research Interests, Lab**
2. **Some Teasers**
3. **Zero-shot [Text Classification | Diagnostic Decision Support]**
4. **Discussion**

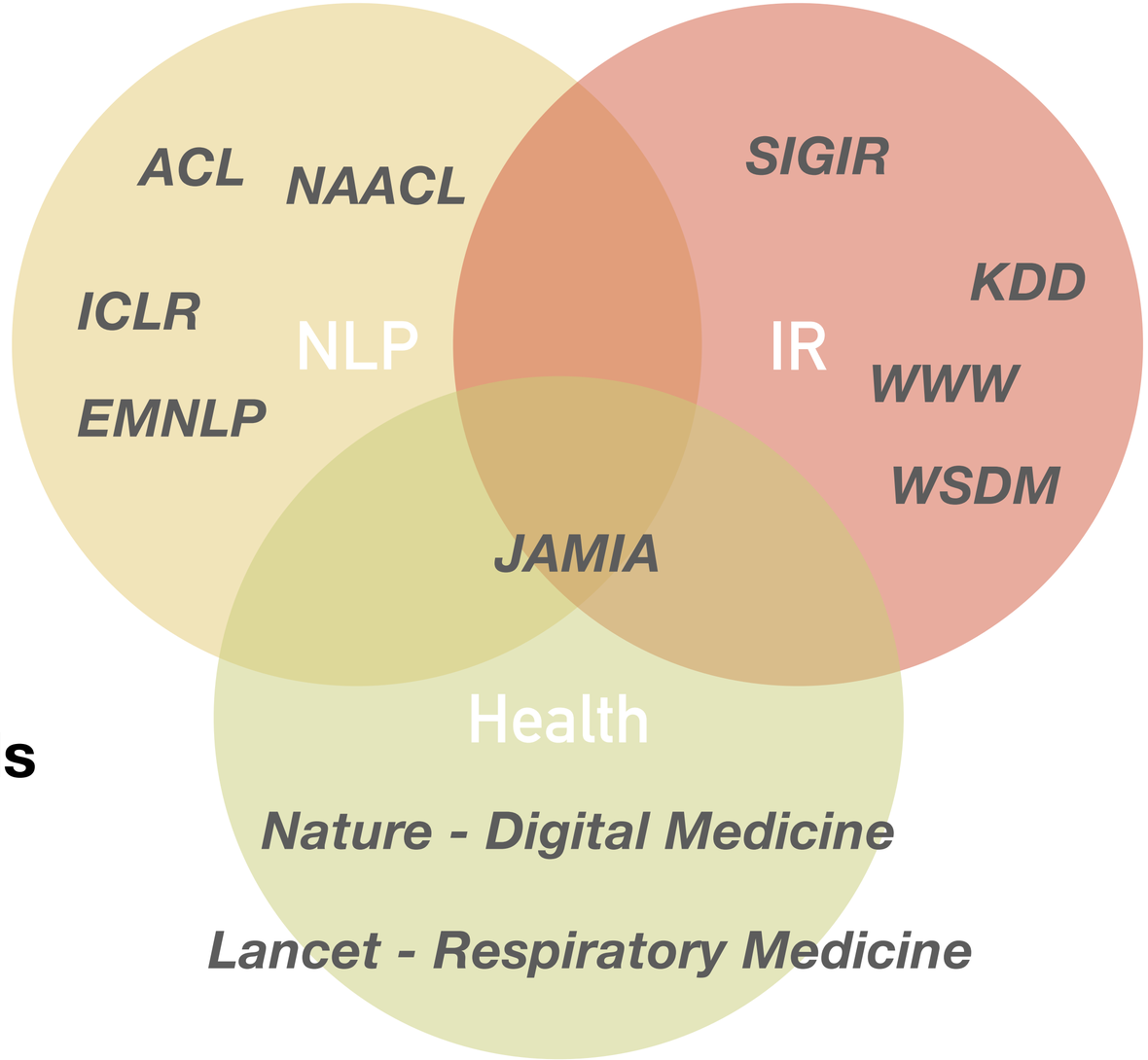
About me

- Hannover
- University of Edinburgh
- TU Delft
- Microsoft
- ETH Zurich
- Harvard University
- Brown University
- University of Tübingen



Interests

- **Dense Retrieval**
- **XIR**
- **Uncertainty Aware Models**
- **Grounded Language Modeling**
- **Manifold Learning for Neural LMs**
- **Clinical Decision Support**



Team

ADEEL ABBASI
ASSISTANT PROFESSOR

ALI BAHRAINIAN
POSTDOC

AMINA ABDULLAHI
PHD STUDENT

CARSTEN EICKHOFF
PROFESSOR, DIRECTOR

CATHERINE CHEN
PHD STUDENT

FLORIAN ROTTACH
PHD STUDENT

GEORGE ZERVEAS
PHD STUDENT

JACK MERULLO
PHD STUDENT

LARS NETH
ADMINISTRATIVE ASSISTANT

LAURA MERCURIO
ASSISTANT PROFESSOR

MARÍA VENEGAS-CARRO
SCIENTIFIC GRANT WRITER

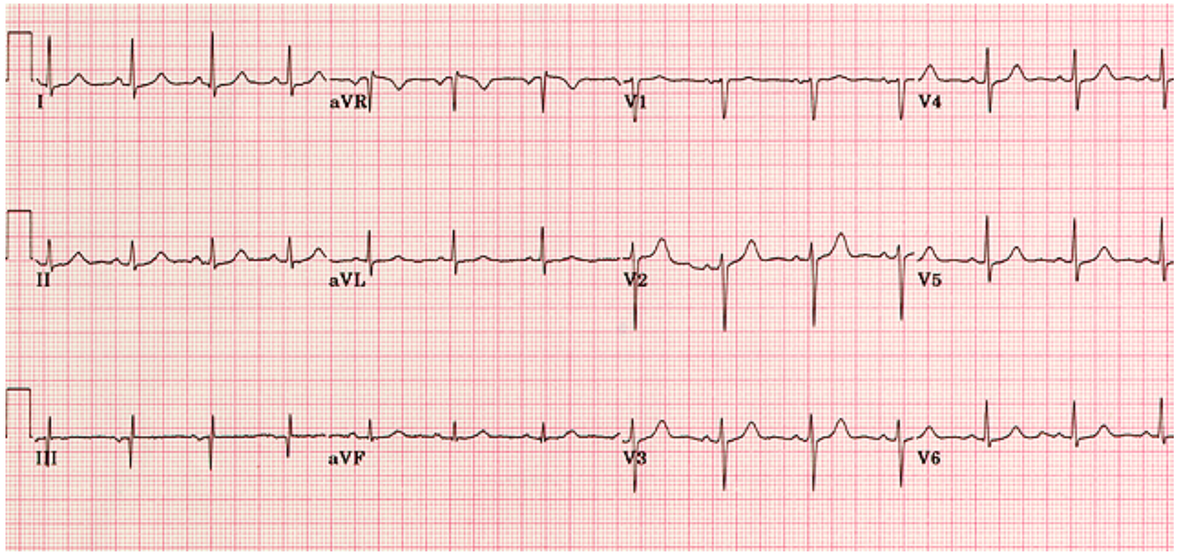
MICHAL GOLOVANEVSKY
PHD STUDENT

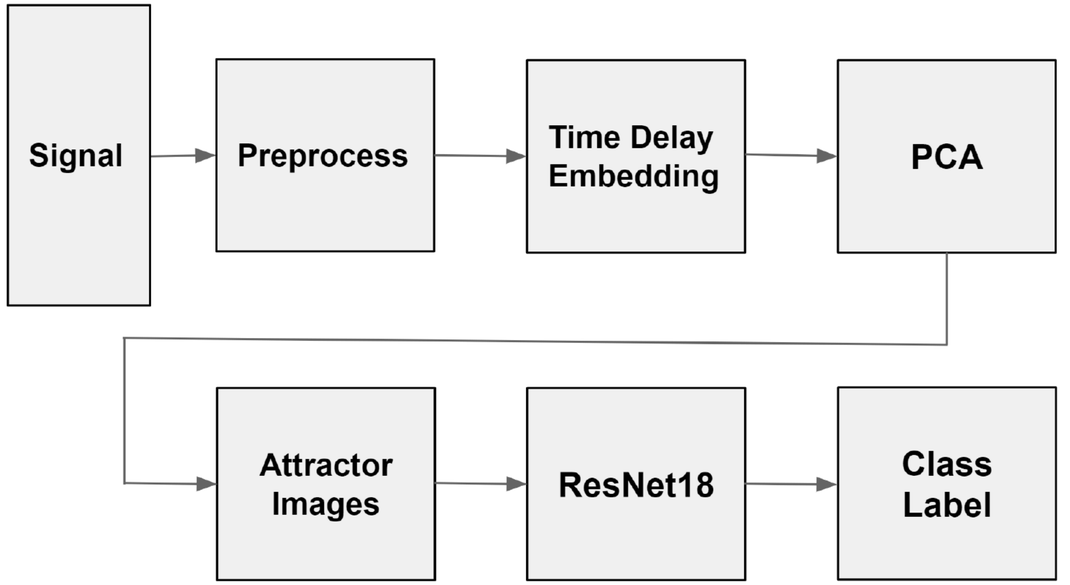
RUOCHEN ZHANG
PHD STUDENT

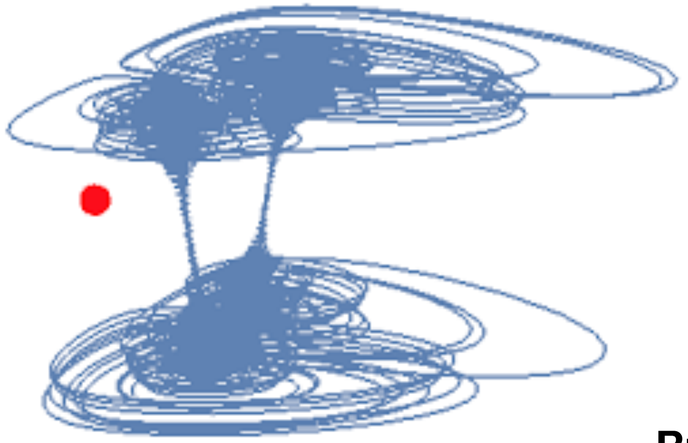
SHAKILA MOSTAAN
SCIENTIFIC GRANT WRITER

WILLIAM RUDMAN
PHD STUDENT

Cardiac Arrhythmia Localization via ECG Attractor Imaging



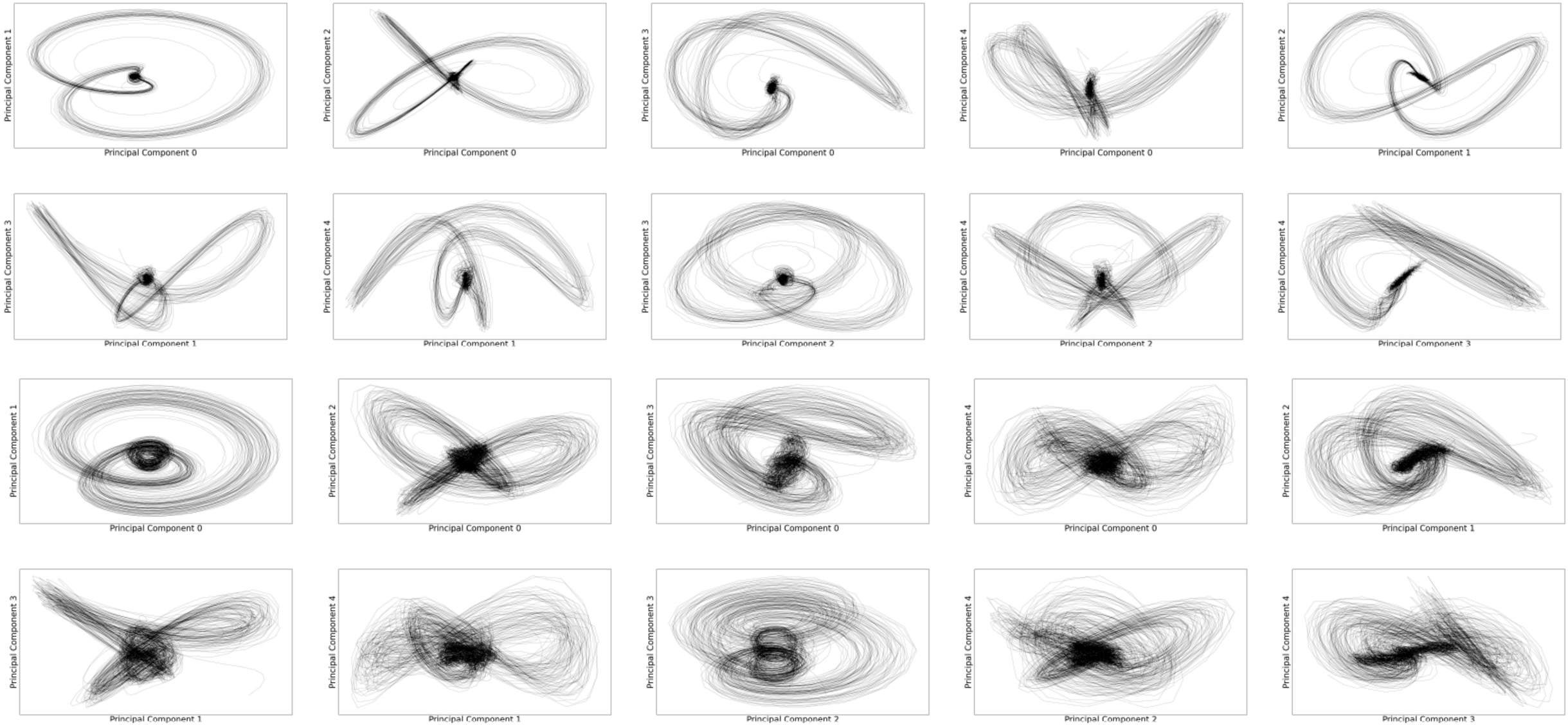




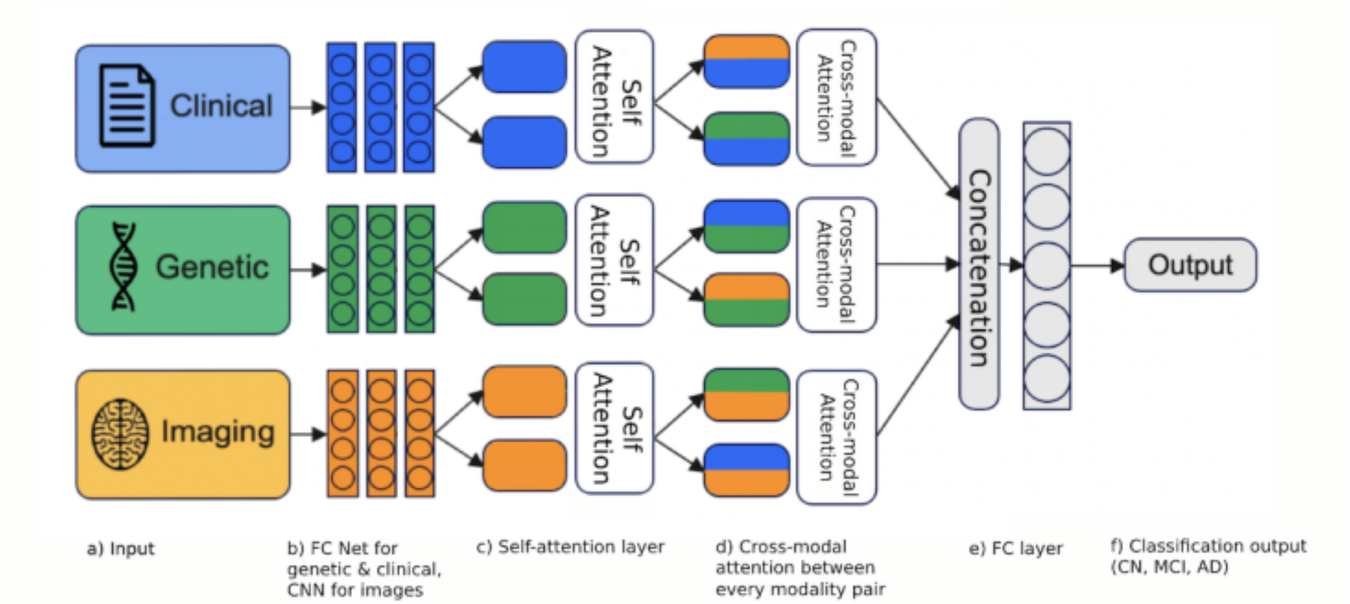
	Normal	AFib	Other	Noise	Avg. Score
Zabihi et. al. [33]	90.49	79.43	75.64	61.11	81.85
Datta et. al. [9]	91.00	79.00	77.00	–	–
Hong et. al. [15]	92.04	86.92	80.68	81.56	85.30
Spectrogram	76.78	43.08	44.71	54.55	54.78
Signal	94.84	96.77	91.93	90.41	93.49
Naïve Attractor	93.93	85.71	94.40	94.87	92.23
01 PCA Attractor	**99.66**	**98.95**	**98.46**	**100**	**99.27**

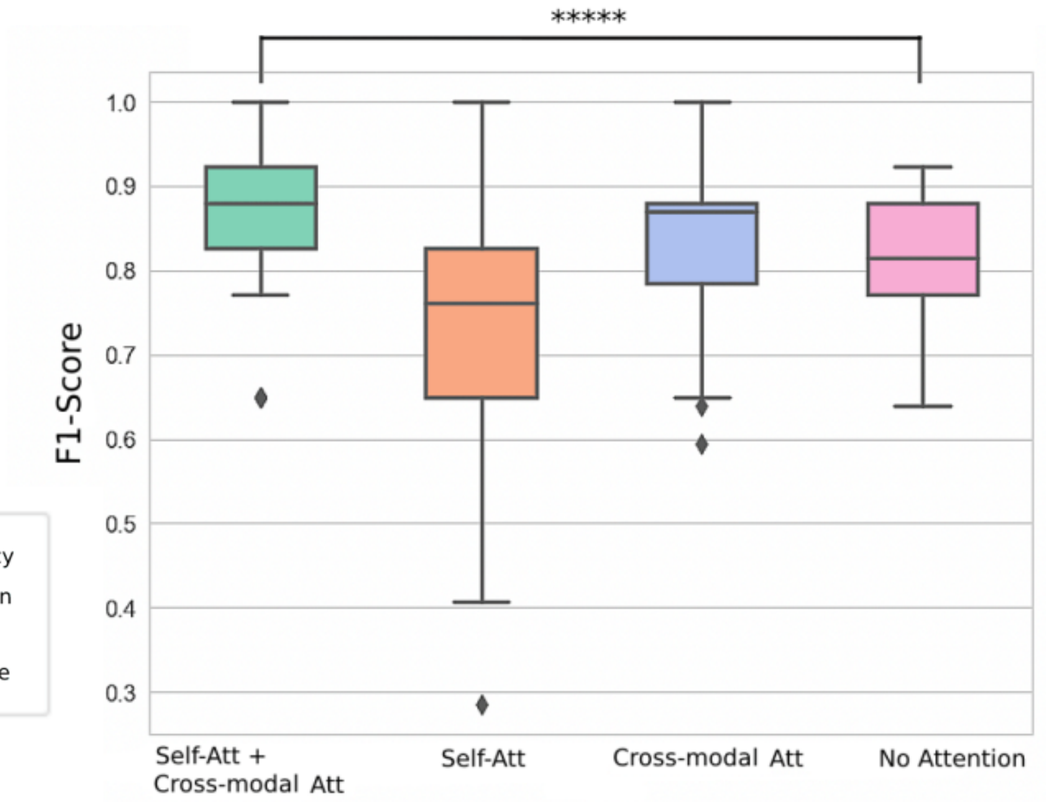
Rudman et al. CinC 2022: ACQuA: Arrhythmia Classification with Quasi-Attractors

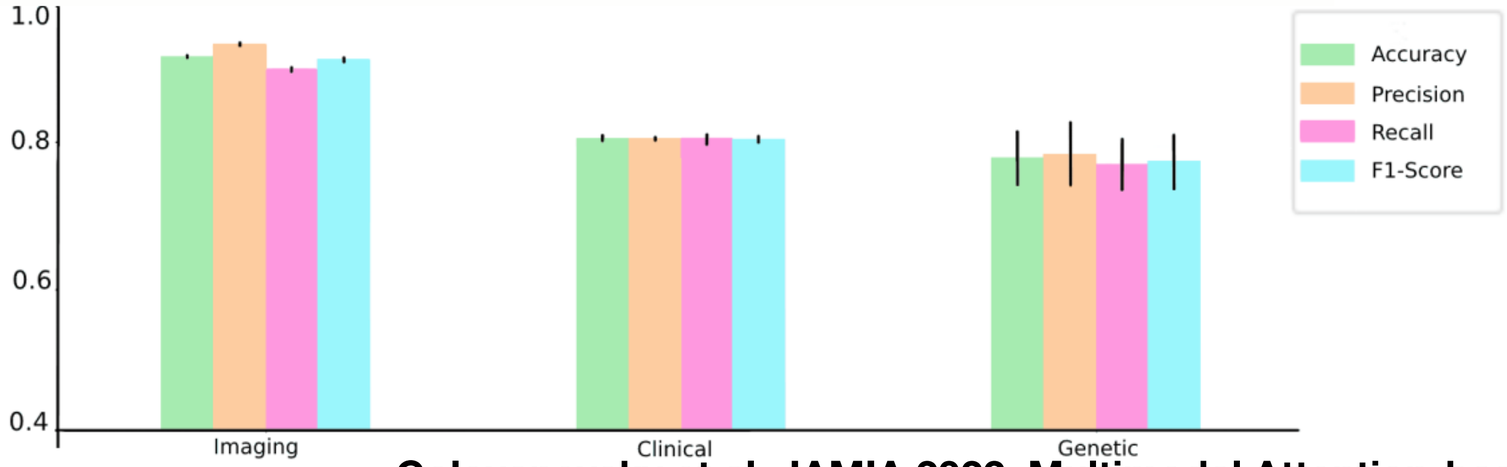
Some Examples



Multimodal Classification

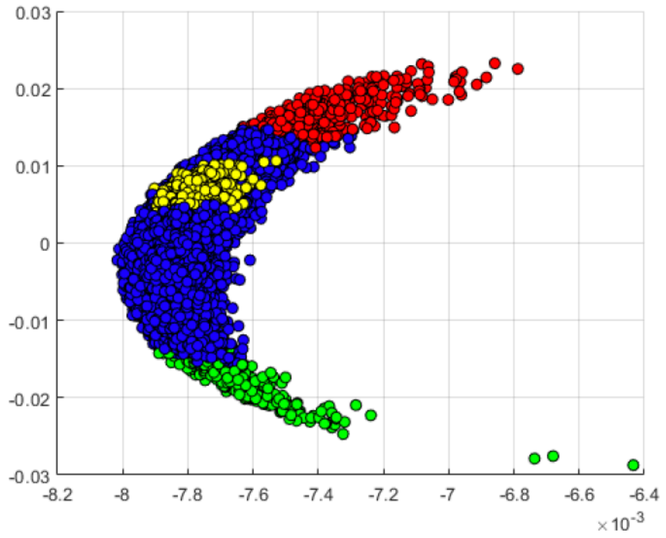


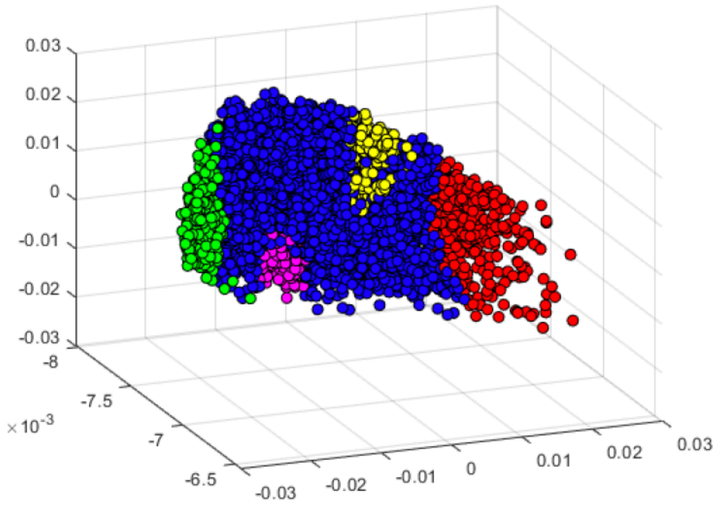




Golovanevsky et al. JAMIA 2022: Multimodal Attention-based Deep Learning for Alzheimer's Disease Diagnosis

Health Outcome Disparities on the EHR





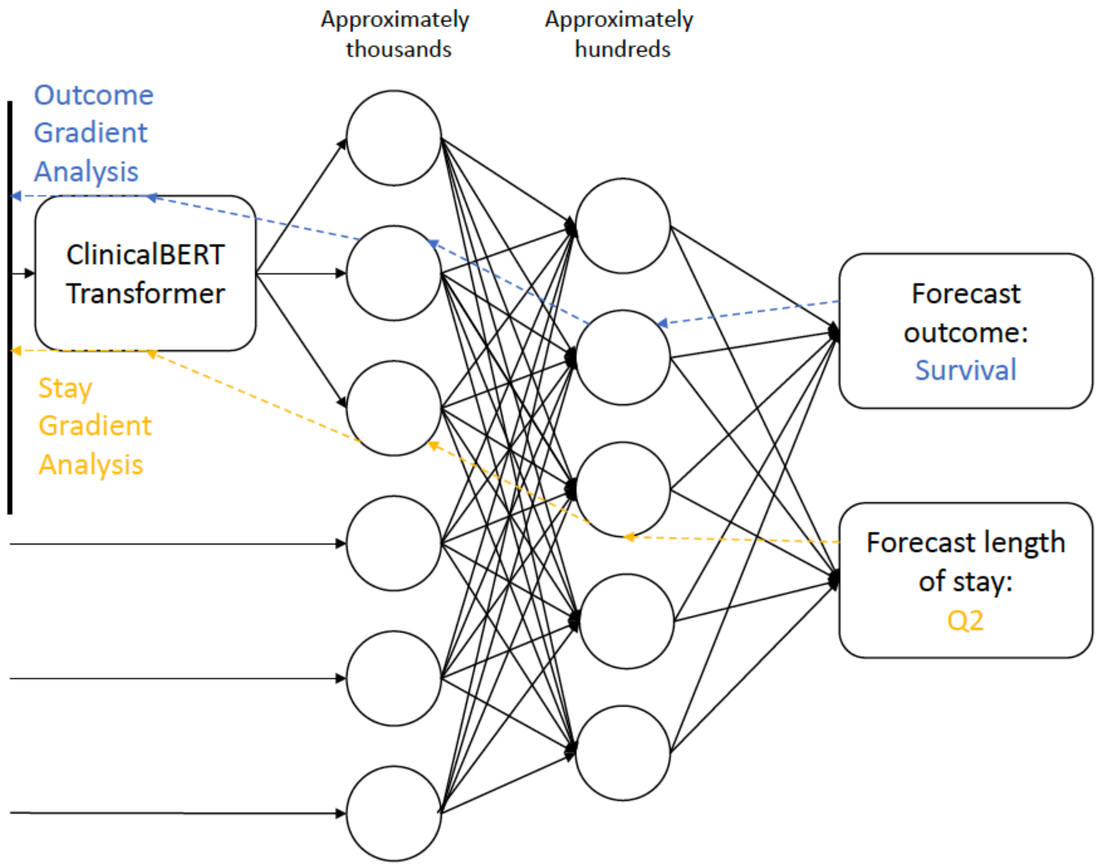
- Self Attention Contextual Embeddings

the patient reported that he had been **feeling well without chest pain,** shortness of breath, or dyspnea on exertion. The patient underwent a cardiac catheterization on the morning of arrival with pci to the native rca and stents and brachytherapy to the vein graft. the patient tolerated the procedure well and approximately **hours later developed a chest pain** noted as out of substernal radiating to his throat and back without shortness of breath, diaphoresis, nausea or vomiting. ekg at that time revealed st elevation in ii, iii, and avf

- Age, sex, insurer…
- Comorbidity 1 yes/no

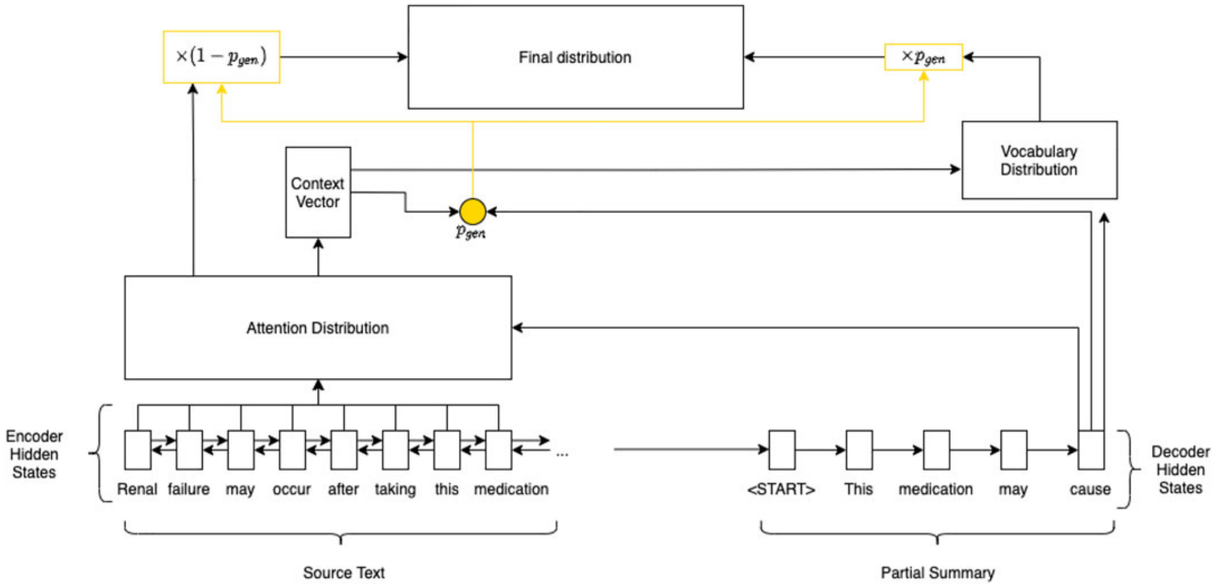
…

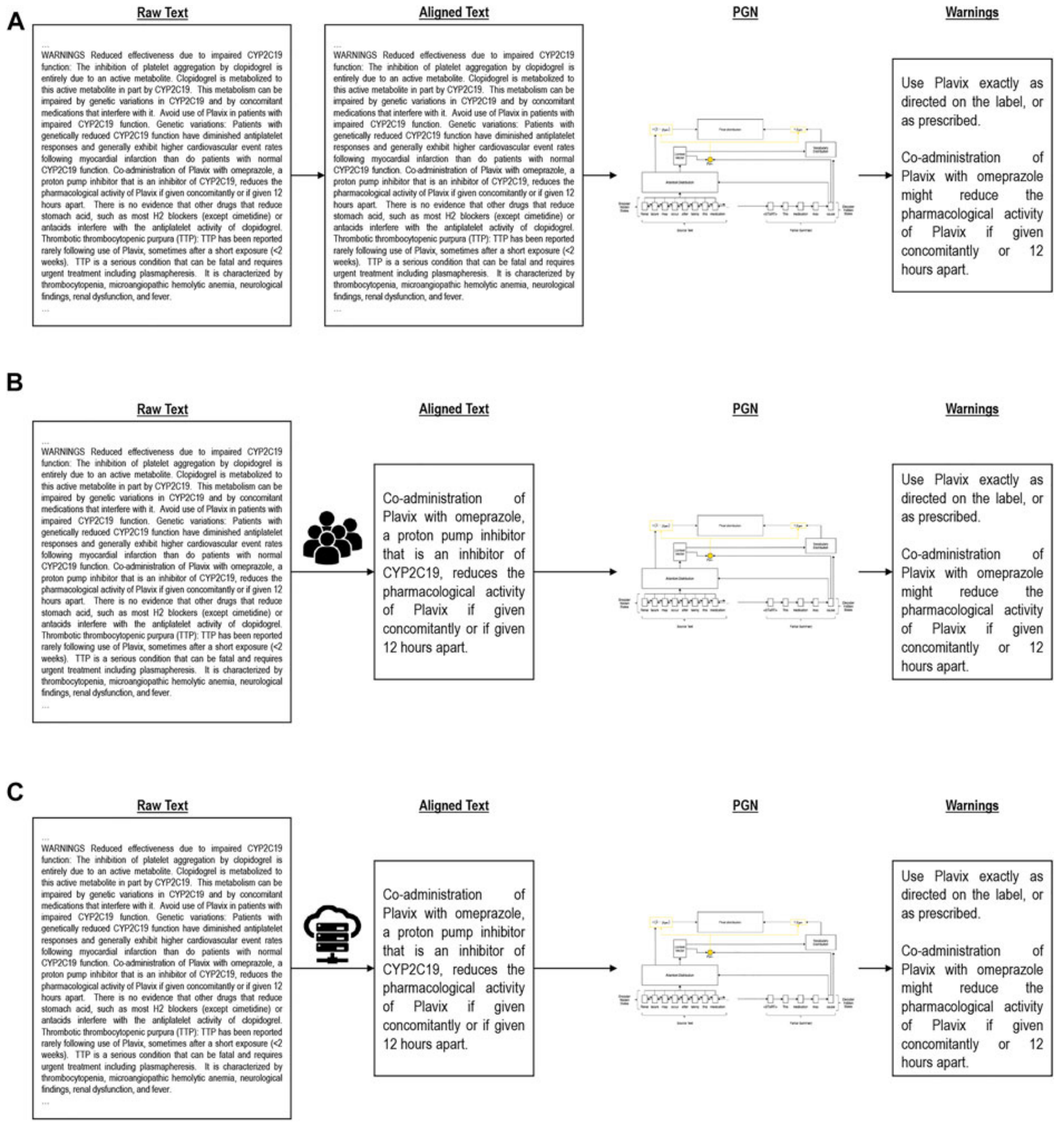
- Comorbidity n yes/no



Garcia-Agundez et al. AMIA 2022: When BERT Fails - The Limits of EHR Classification

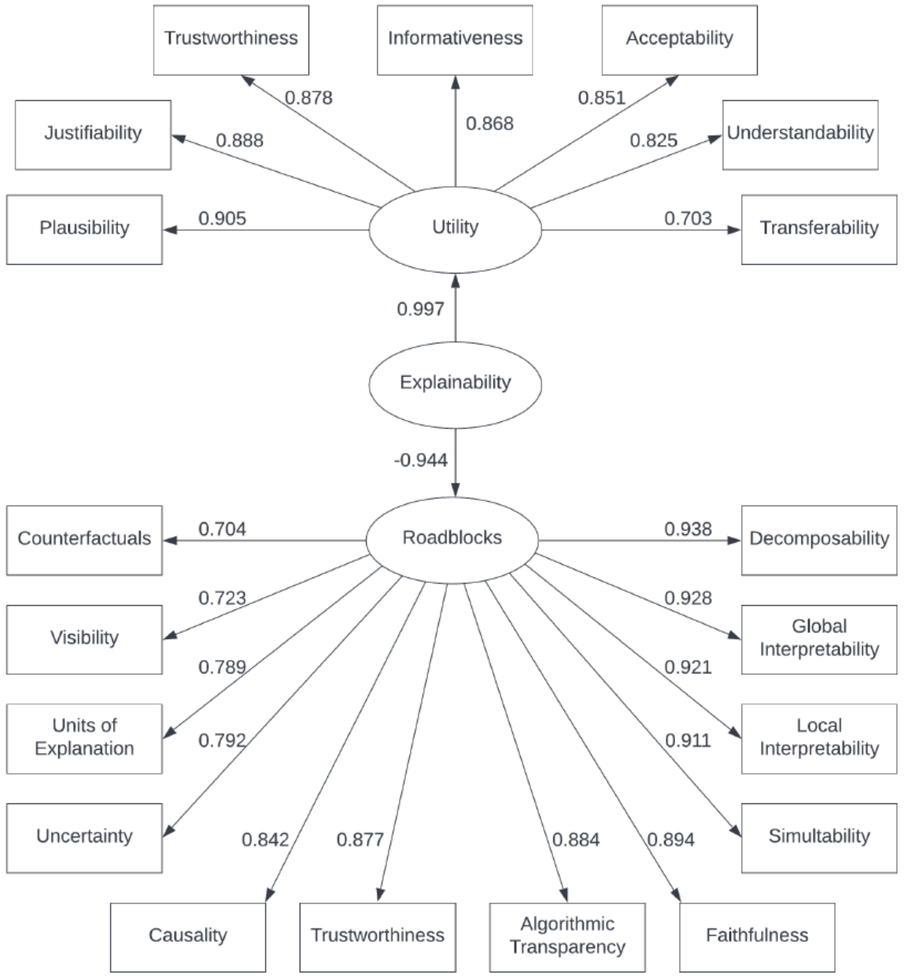
Generating Drug Leaflets





Meyer et al. Frontiers in Pharmacology 2023: Neural text Generation in Regulatory Medical Writing

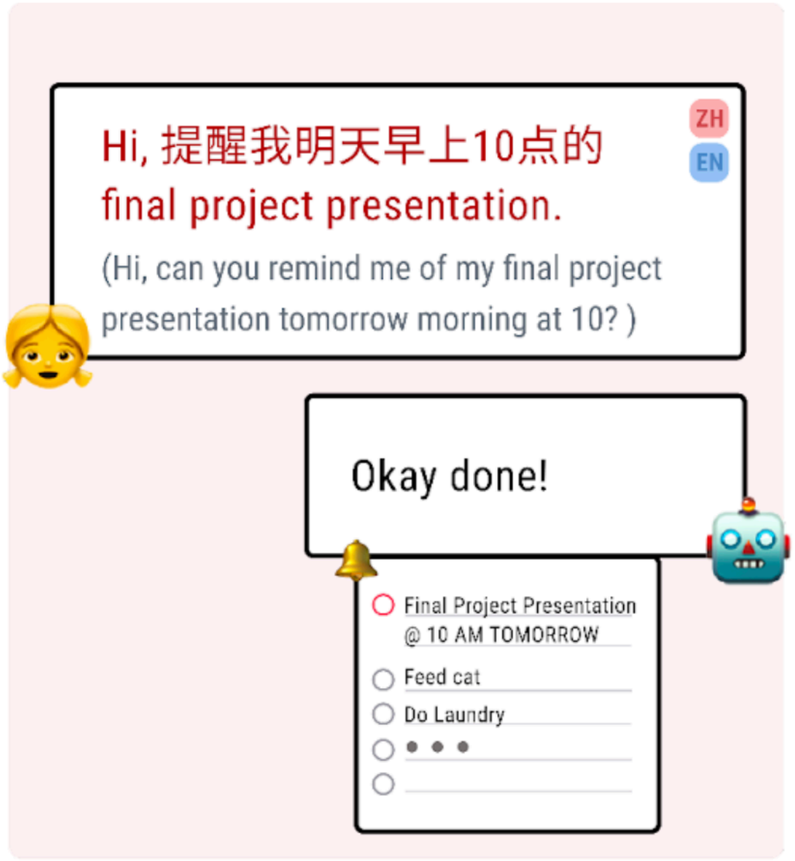
Measuring System Explainability

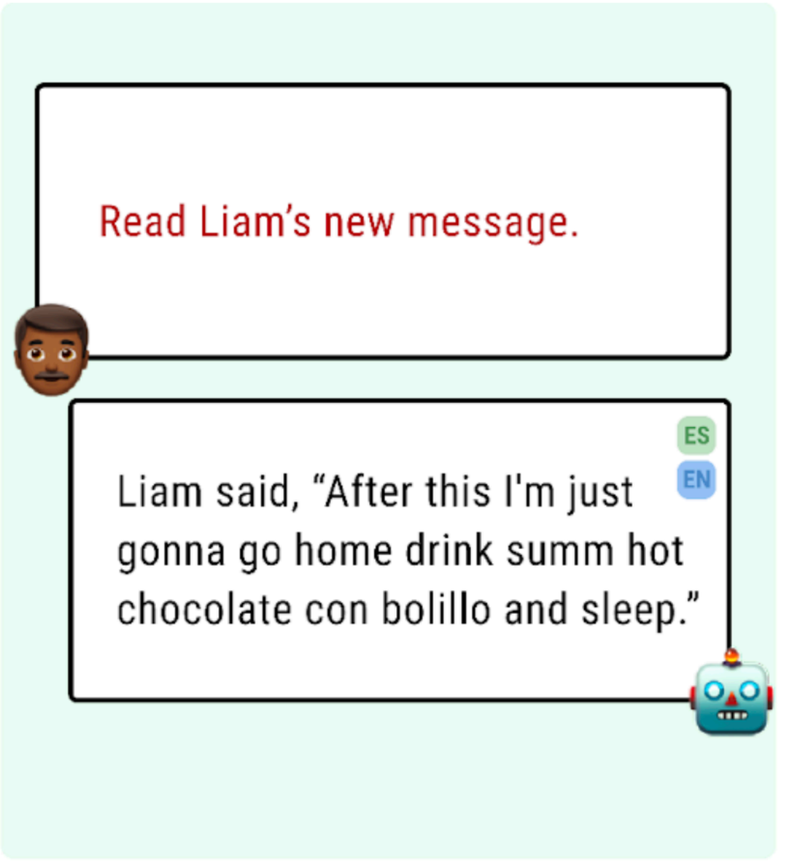


Factors & Items	Original Factor Label	Std. Loading	Std. Error
Factor 1		0.997	0.000
41	Plausibility	0.905	0.096
49	Justifiability	0.888	0.097
47	Trustworthiness	0.878	0.096
21	Informativeness	0.868	0.094
37	Acceptability	0.851	0.110
19	Understandability	0.825	0.087
15	Transferability	0.703	0.000
Factor 2		-0.944	0.065
2	Decomposability	0.938	0.094
22	Global Interpretability	0.928	0.098
24	Local Interpretability	0.921	0.098
0	Simultability	0.911	0.099
38	Faithfulness	0.894	0.097
4	Algorithmic Transparency	0.884	0.097
46	Trustworthiness	0.877	0.091
6	Causaility	0.842	0.095
8	Uncertainty	0.792	0.092
34	Units of Explanation	0.789	0.094
12	Visibility	0.723	0.092
26	Counterfactuals	0.704	0.000

Chen et al. *under review*: Evaluating Search Explainability with Psychometrics and Crowdsourcing

Language Generation in a Code-switched World





Zhang et al. *under review*: CroCoSum A Benchmark Dataset for Cross-Lingual Code-Switched Summarization

Are Language Models World Models?

A frozen image encoder encodes an image as a feature map

A linear projection is tuned to project from image space to text space

The image projections are fed as soft prompts into a generative LM

Image Encoder → Linear Proj. → Text Decoder → "A picture of a dog on a skateboard"

Image Captioning

CLIP	a giraffe in the lobby of the building
NFRN50	the giraffe in the zoo.
BEIT	a peacock in the garden
NFRN50 Random	a man and a woman in a field of flowers

CLIP	tennis player in action
NFRN50	tennis player at the tennis tournament.
BEIT	tennis player during a tennis match.
NFRN50 Random	the new logo for the team

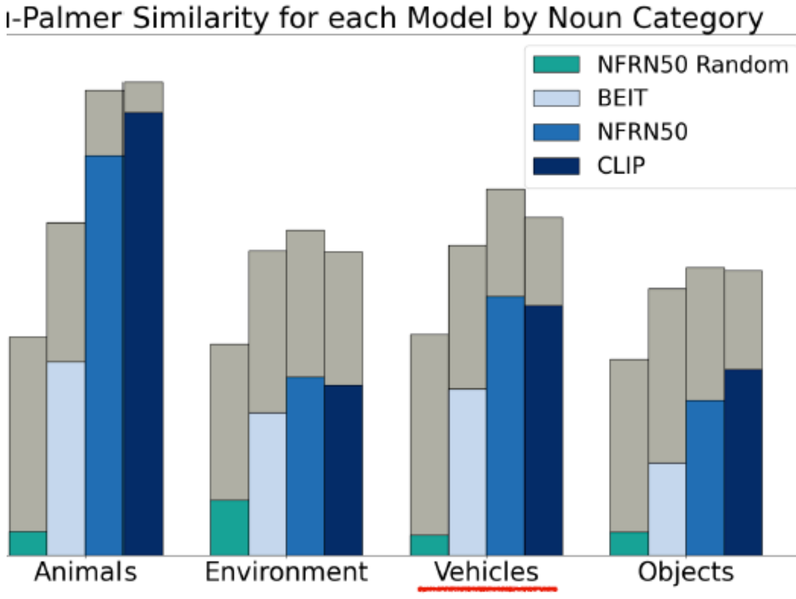
Visual Question Answering

CLIP	He is surfing a wave.
NFRN50	He is surfing the waves.
BEIT	He is jumping into the water.
NFRN50 Random	He is swimming in the pool.

Q: What is the person doing?
A: surfing

CLIP	A tennis racket
NFRN50	A tennis racket
BEIT	A baseball bat.
NFRN50 Random	A tree

Q: What is the person holding?
A: tennis racket

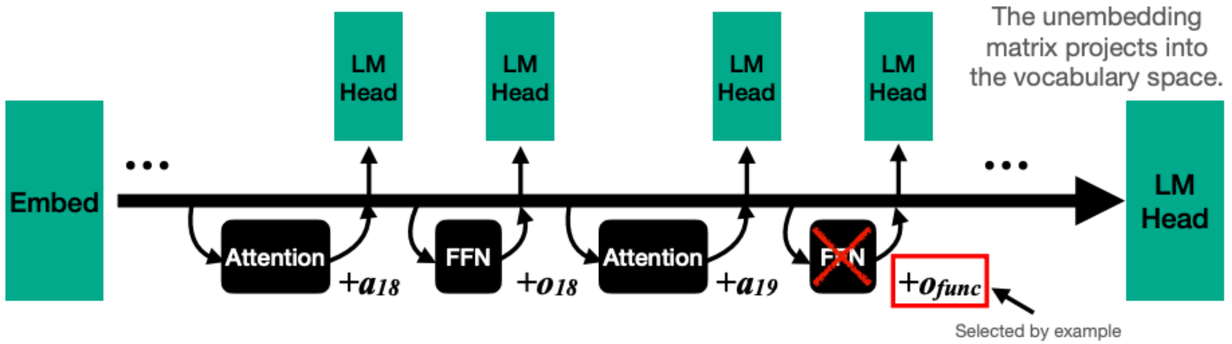


Merullo et al. ICLR 2023: Linearly Mapping from Image to Text Space

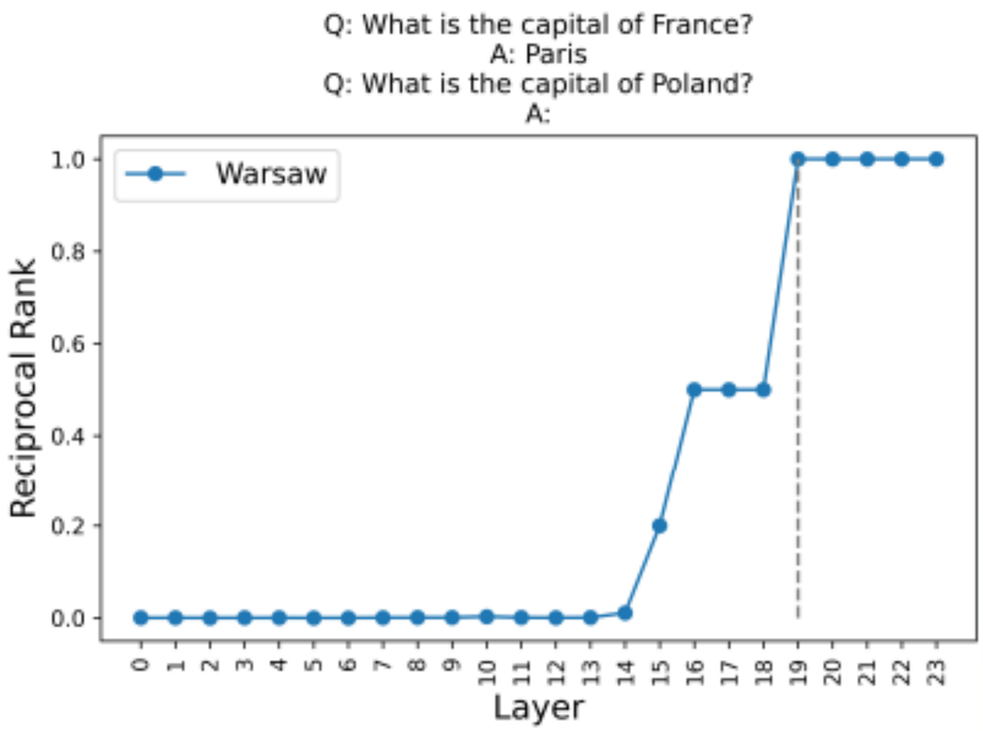
Grounding Language Models in Physics

Merullo et al. *SEM 2022: Pretraining on Interactions for Learning Grounded Affordance Representations

LLM Vector Arithmetics

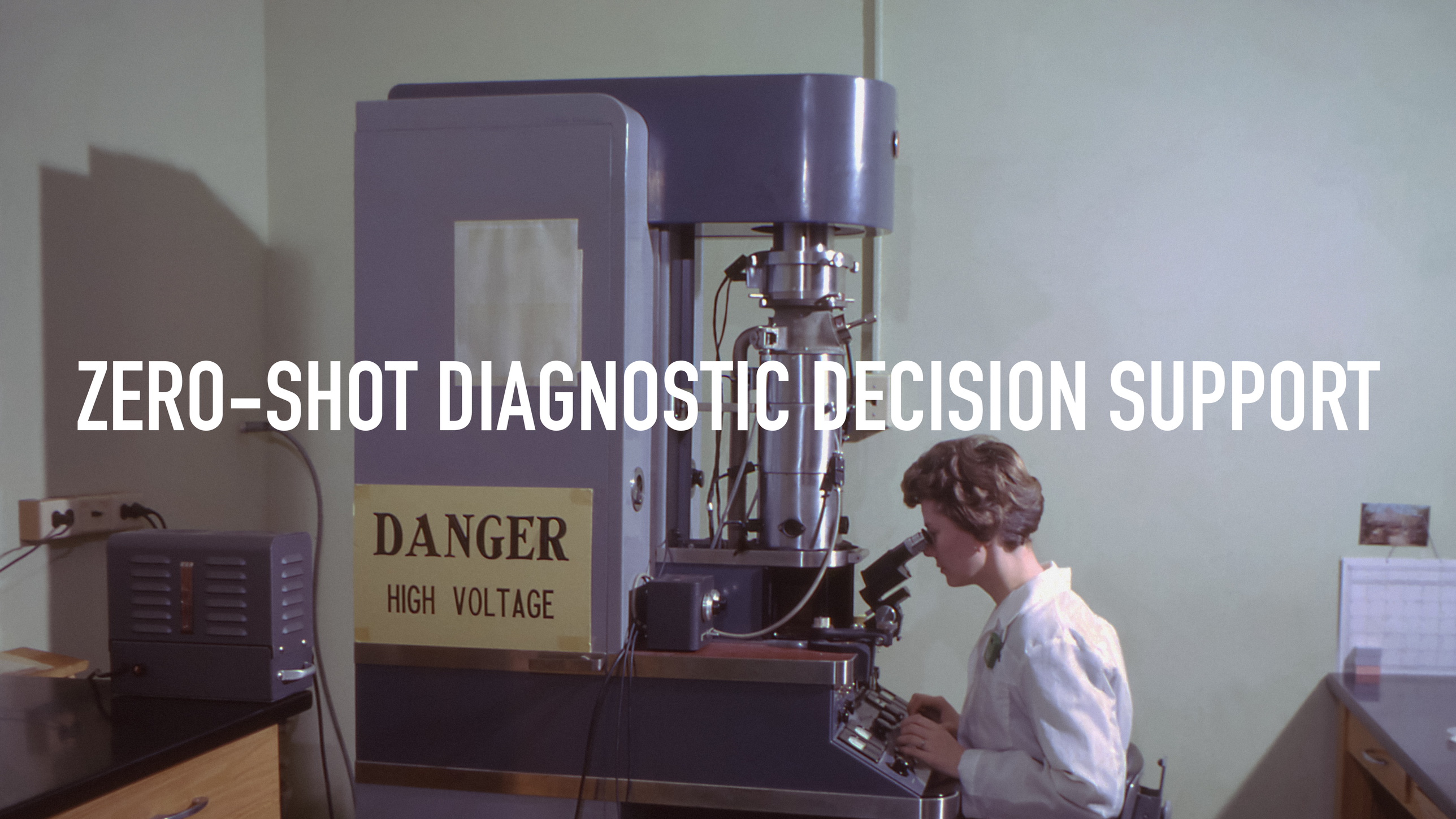


- ***We find simple vector arithmetic in Transformers***
- ***3 Models***
  - ***GPT2 (124M)***
  - ***GPT-J (6B)***
  - ***Bloom (176B)***

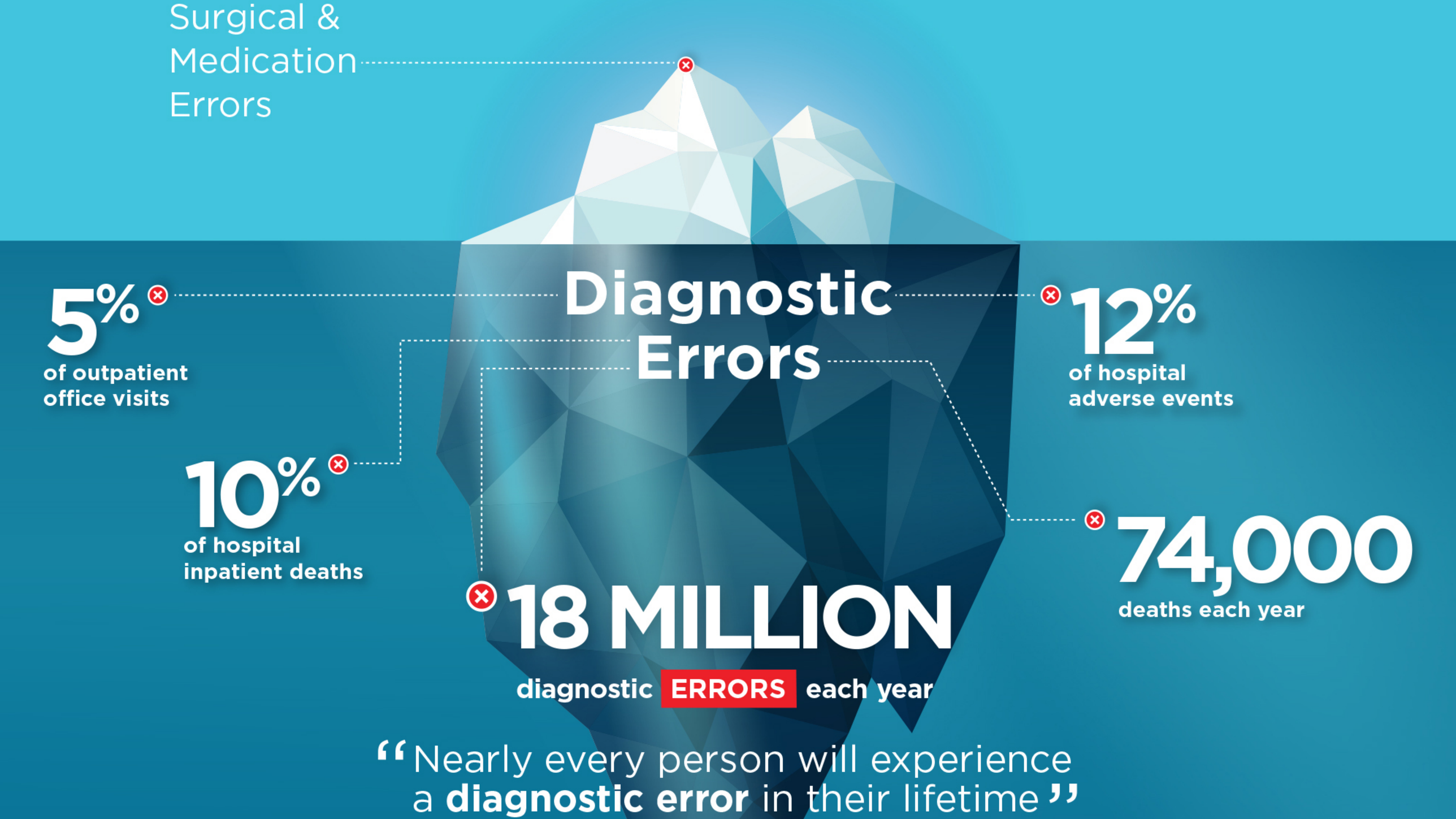


Layer	Top Token	
0	(	
1	A	
2	A	
3	A	
4	A	
5	A	
6	No	
7	C	
8	A	
9	A	
10	A	
11	A	
12	Unknown	
13	C	
14	St	
15	Poland	A
16	Poland	A
17	Poland	A
18	Poland	A
19	Warsaw	B
20	Warsaw	C
21	Warsaw	C
22	Warsaw	C
23	Warsaw	C

Merullo et al. *under review*: Language Models Implement Simple Word2Vec-style Vector Arithmetic



ZERO-SHOT DIAGNOSTIC DECISION SUPPORT


Surgical &
Medication
Errors
Diagnostic
Errors
5%
of outpatient
office visits
10%
of hospital
inpatient deaths
18 MILLION
diagnostic ERRORS each year
12%
of hospital
adverse events
74,000
deaths each year
“Nearly every person will experience
a diagnostic error in their lifetime”

ML to the Rescue

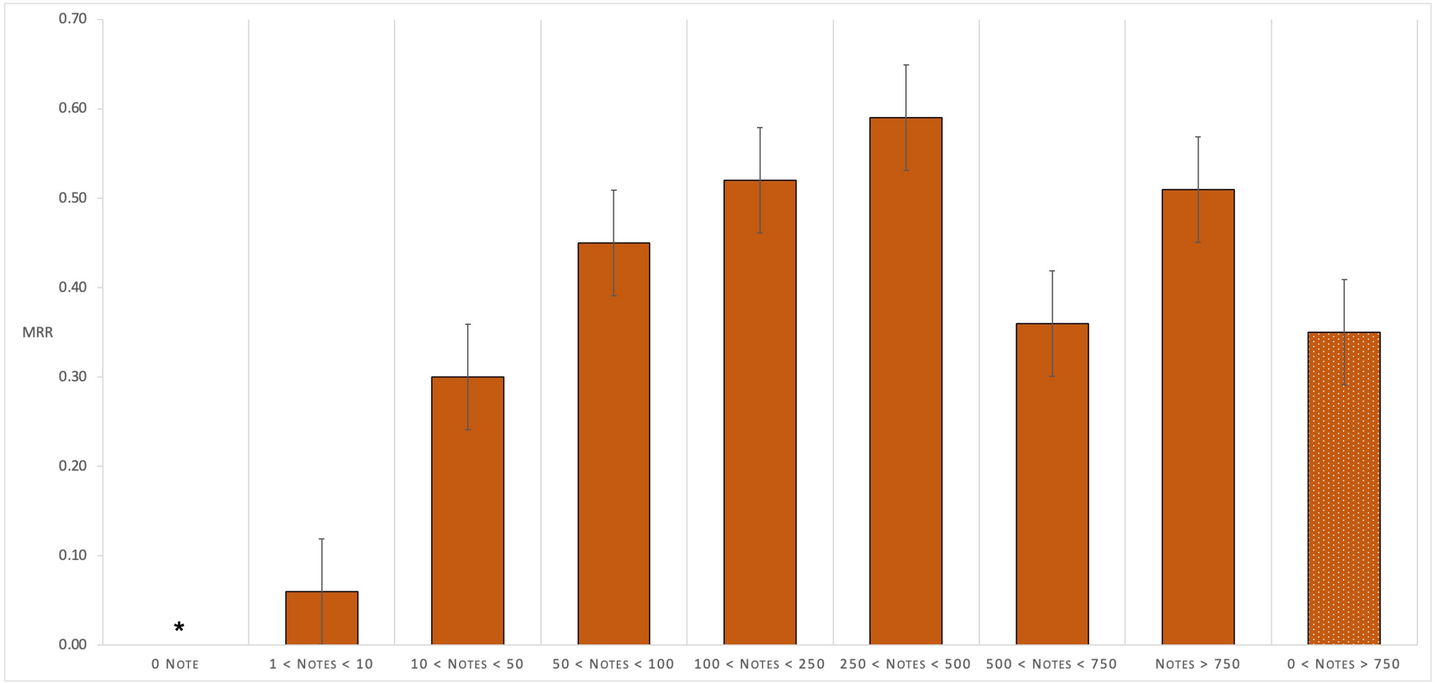
- *Golovanevsky et al. 2022*: Alzheimers (92.28%)
- Delahanty et al. 2018: Sepsis (97%)
- *Gulshan et al. 2016*: Diabetic Retinopathy (99.1%)
- *Rudman et al. 2022:* Cardiac Arrhythmias (99.27%)
- …

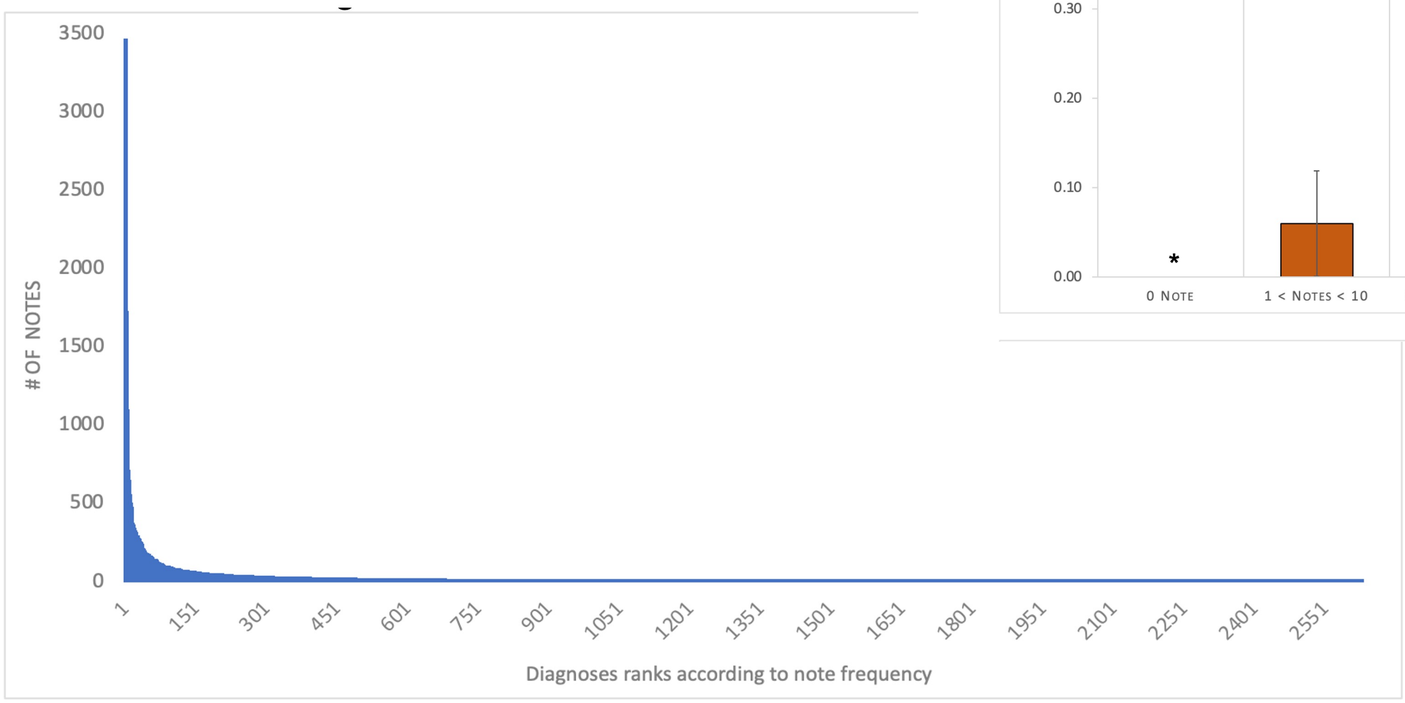
ML to the Rescue

- *Golovanevsky et al. 2022*: Alzheimers (92.28%) [n= 2,384]
- Delahanty et al. 2018: Sepsis (97%) [n= 2,759,529]
- *Gulshan et al. 2016*: Diabetic Retinopathy (99.1%) [n= 128,175]
- *Rudman et al. 2022:* Cardiac Arrhythmias (99.27%) [n= 8,528]
- …

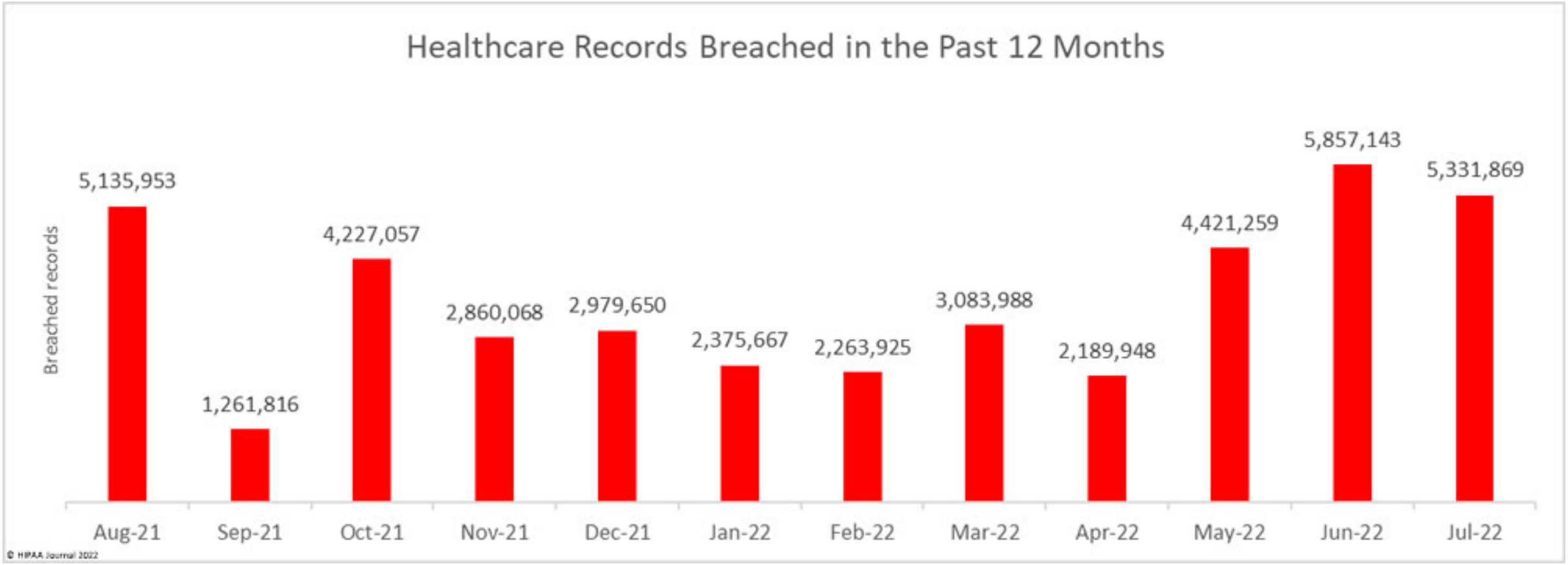
But only if you have the Data!

A.

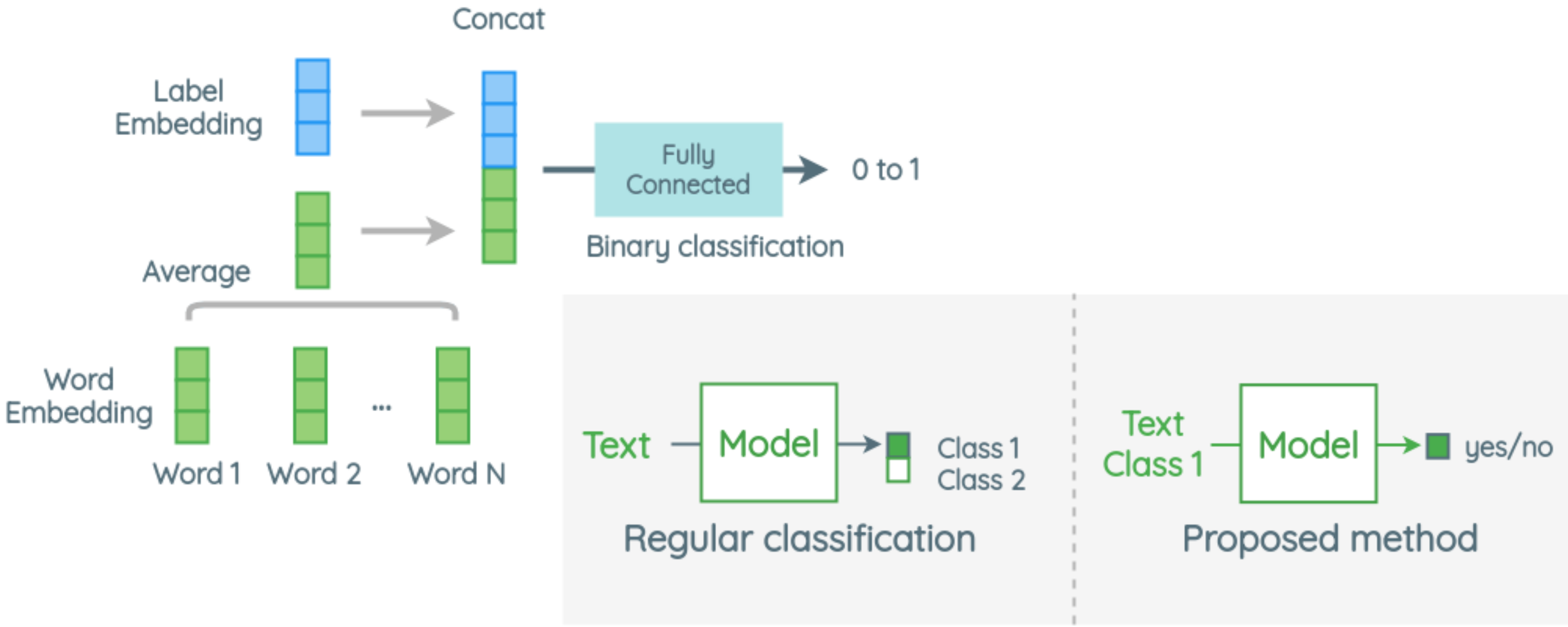




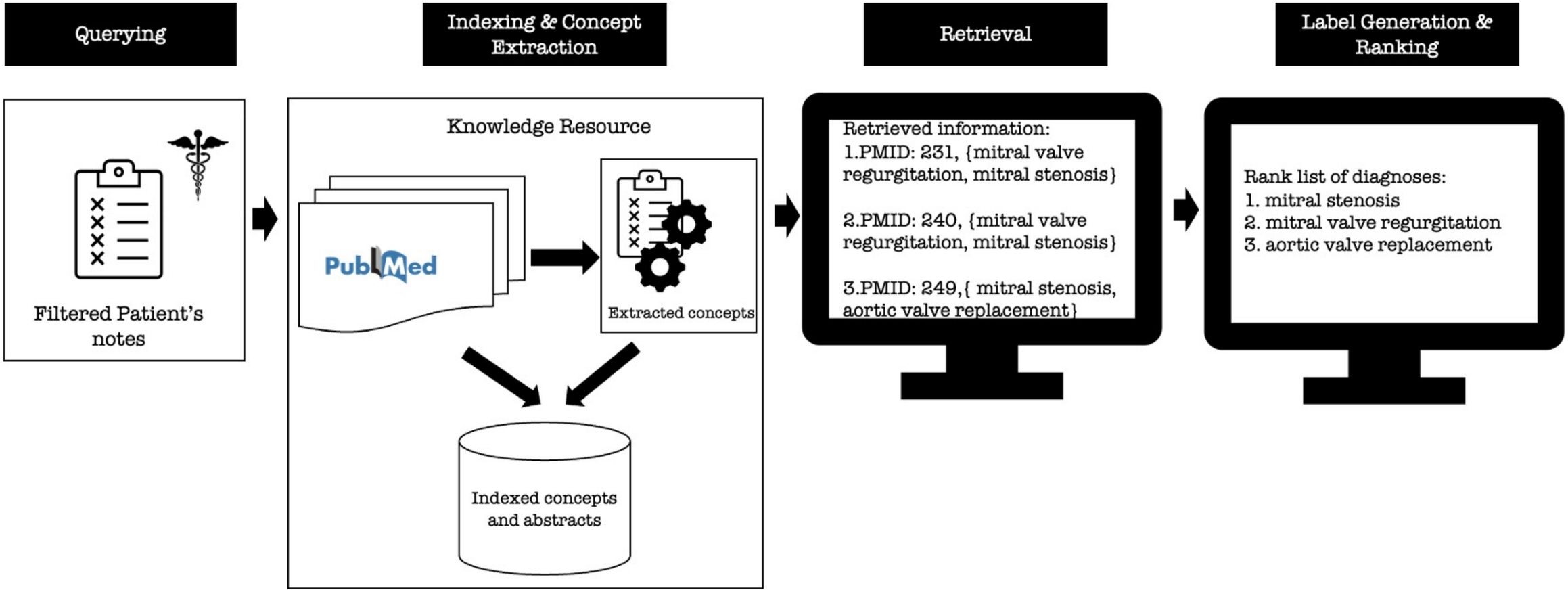
Sharing is …


Healthcare Records Breached in the Past 12 Months
Breached records
5,135,953
1,261,816
4,227,057
2,860,068
2,979,650
2,375,667
2,263,925
3,083,988
2,189,948
4,421,259
5,857,143
5,331,869
Aug-21
Sep-21
Oct-21
Nov-21
Dec-21
Jan-22
Feb-22
Mar-22
Apr-22
May-22
Jun-22
Jul-22
© HIPAA Journal 2022

Zero-shot Text Classification



Via Retrieval

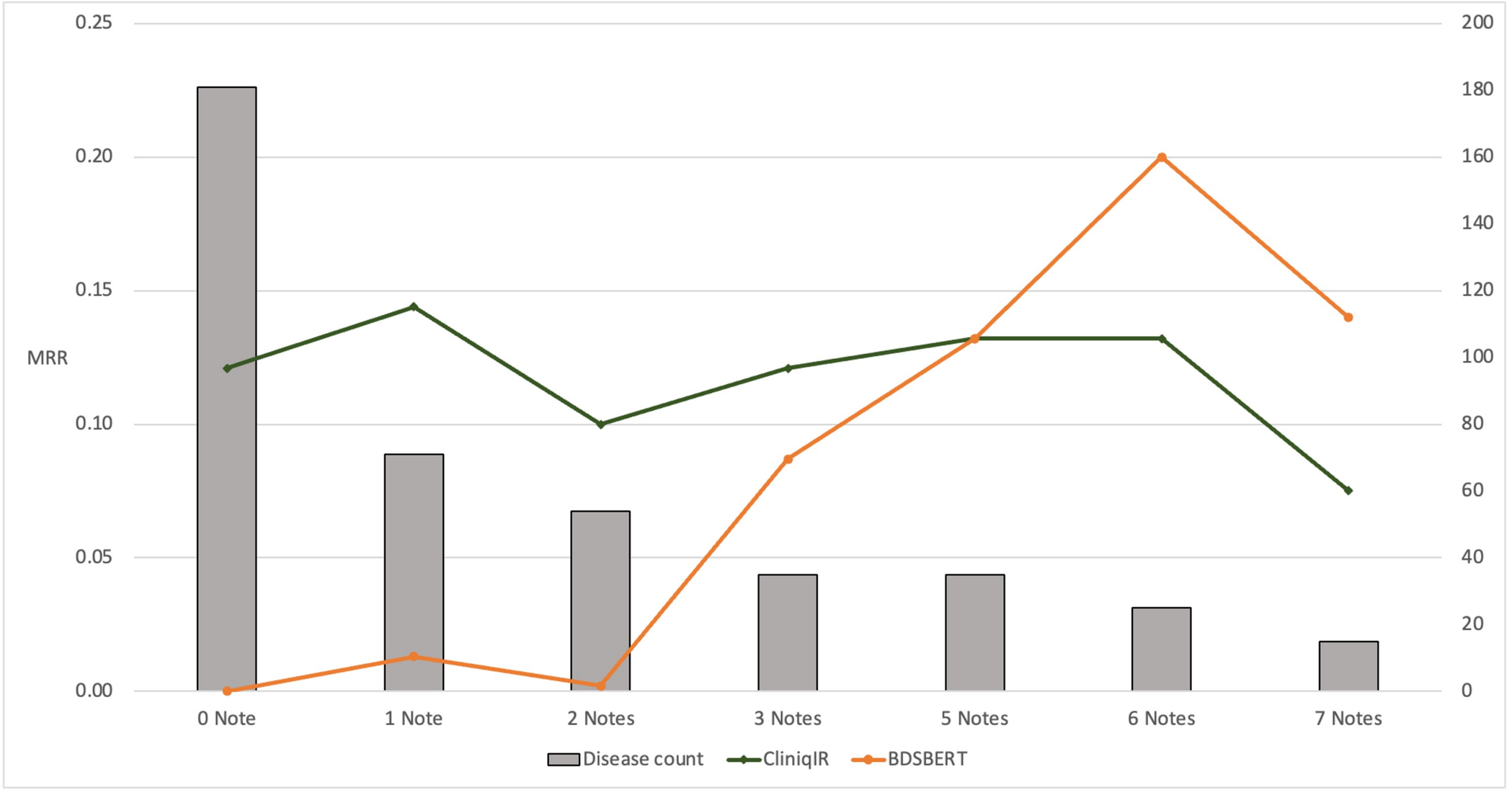


Datasets

- PubMed
  - 33M abstracts
- MIMIC III
  - 50,000 ICU encounters
  - 2643 unique diagnoses
  - 902 of them occur exactly once
- DC3
  - 31 (difficult) case reports

Results I

A.


0.25
0.20
0.15
0.10
0.05
0.00
MRR
200
180
160
140
120
100
80
60
40
20
0
0 Note
1 Note
2 Notes
3 Notes
5 Notes
6 Notes
7 Notes
Disease count
CliniqIR
BDSBERT

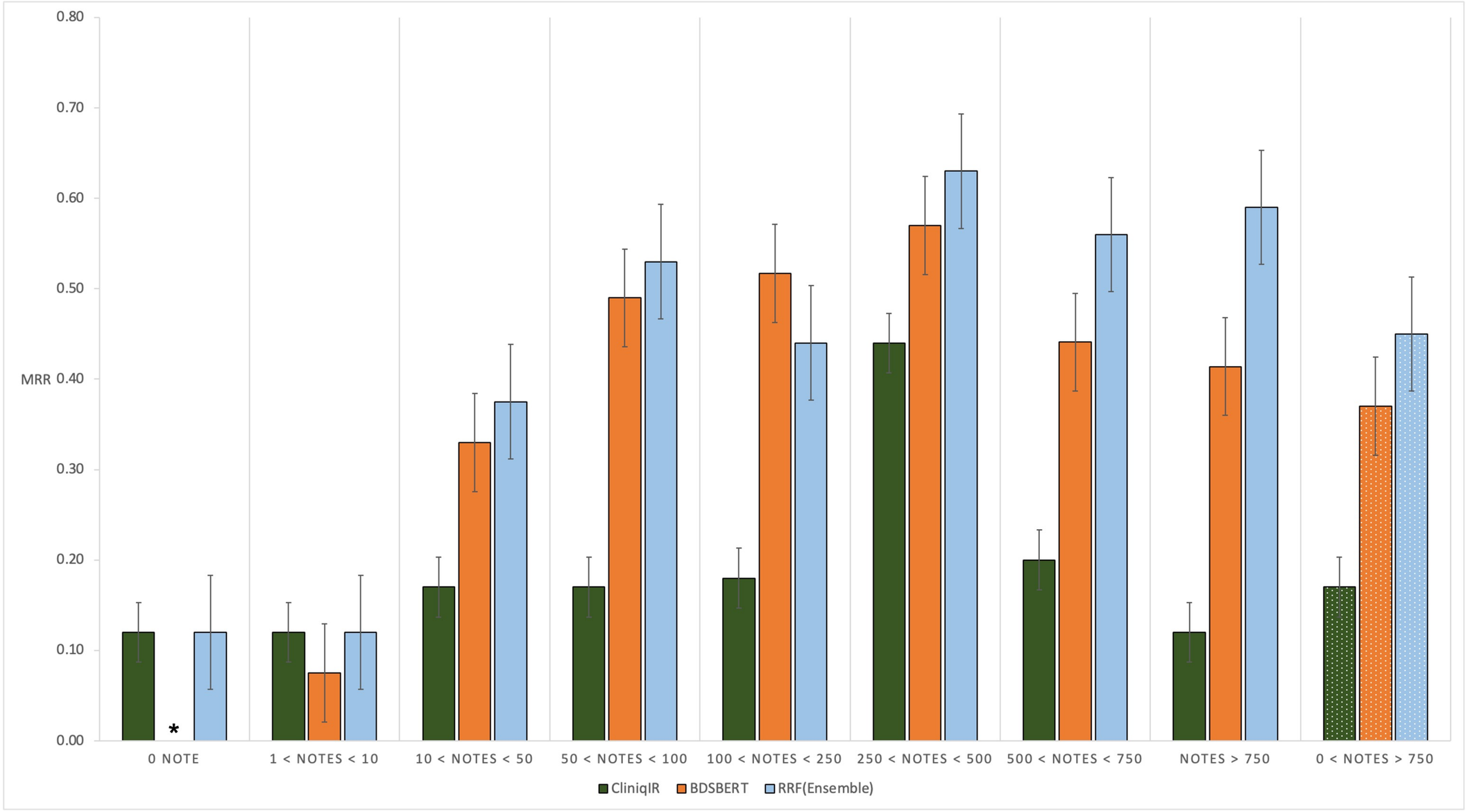
Ensembling

$$P'(d,e) = \frac{P_s(d,e) + \mu P_u(d,e)}{|d| + \mu}$$

- d: diagnosis
- e: encounter
- μ: Dirichlet prior, median # of training examples per diagnosis
- |d|: # of training examples for diagnosis d

Results II

B.


0.80
0.70
0.60
0.50
MRR 0.40
0.30
0.20
0.10
0.00
*
0 NOTE
1 < NOTES < 10
10 < NOTES < 50
50 < NOTES < 100
100 < NOTES < 250
250 < NOTES < 500
500 < NOTES < 750
NOTES > 750
0 < NOTES > 750
CliniqIR
BDSBERT
RRF(Ensemble)

Reasons to Care

- The world is Zipfian
- Effective zero-shot approaches rely on structured information (KB triples)
- Unstructured data is abundantly available and growing fast
- Using unstructured collections for unsupervised learning unlocks considerable resources
- Next stop: Going beyond diagnostics


DISCUSSION