Language Models in Complex Diagnostics

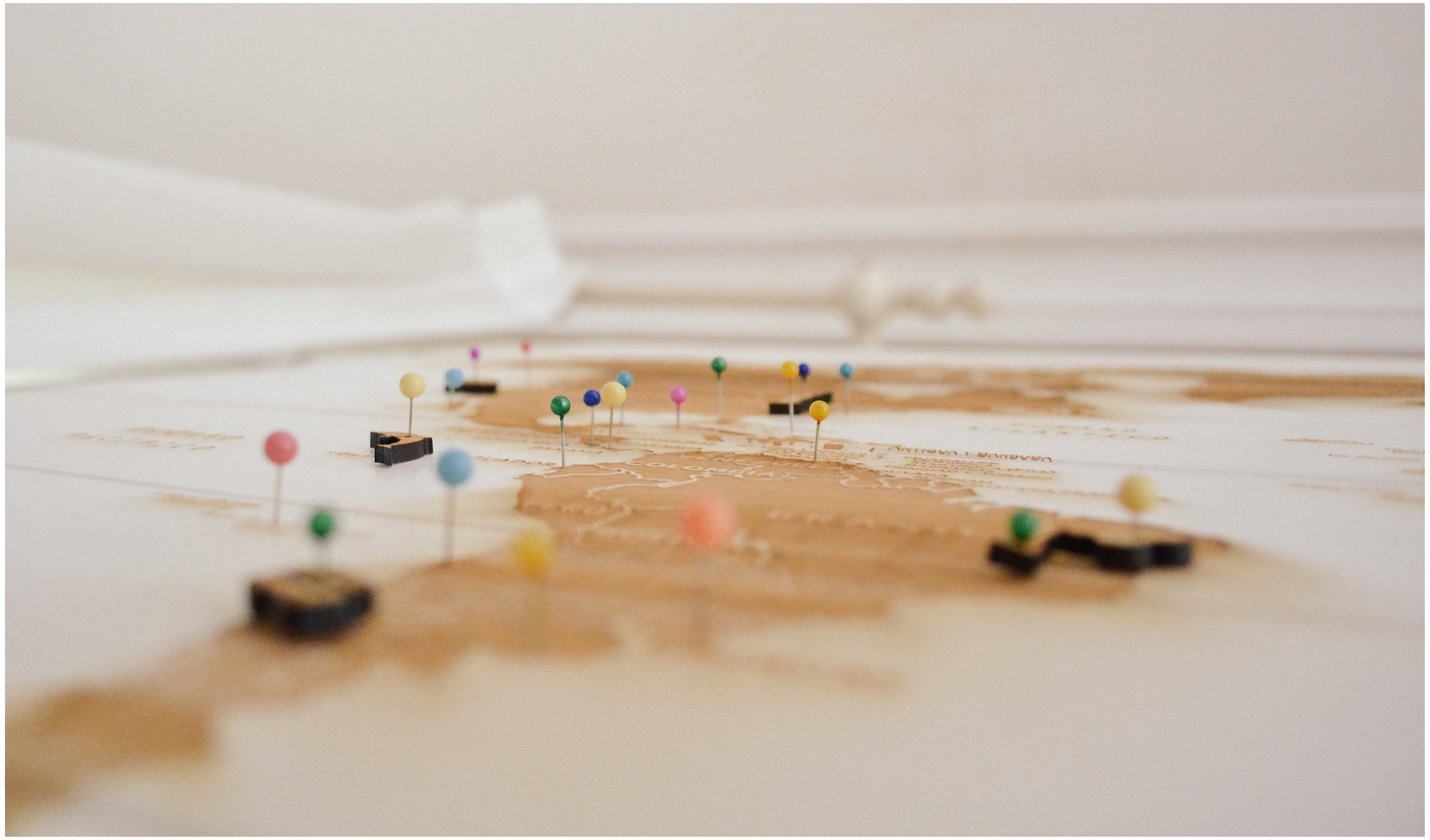
Carsten Eickhoff

EBERHARD KARLS
UNIVERSITÄT
TÜBINGEN

BROWN

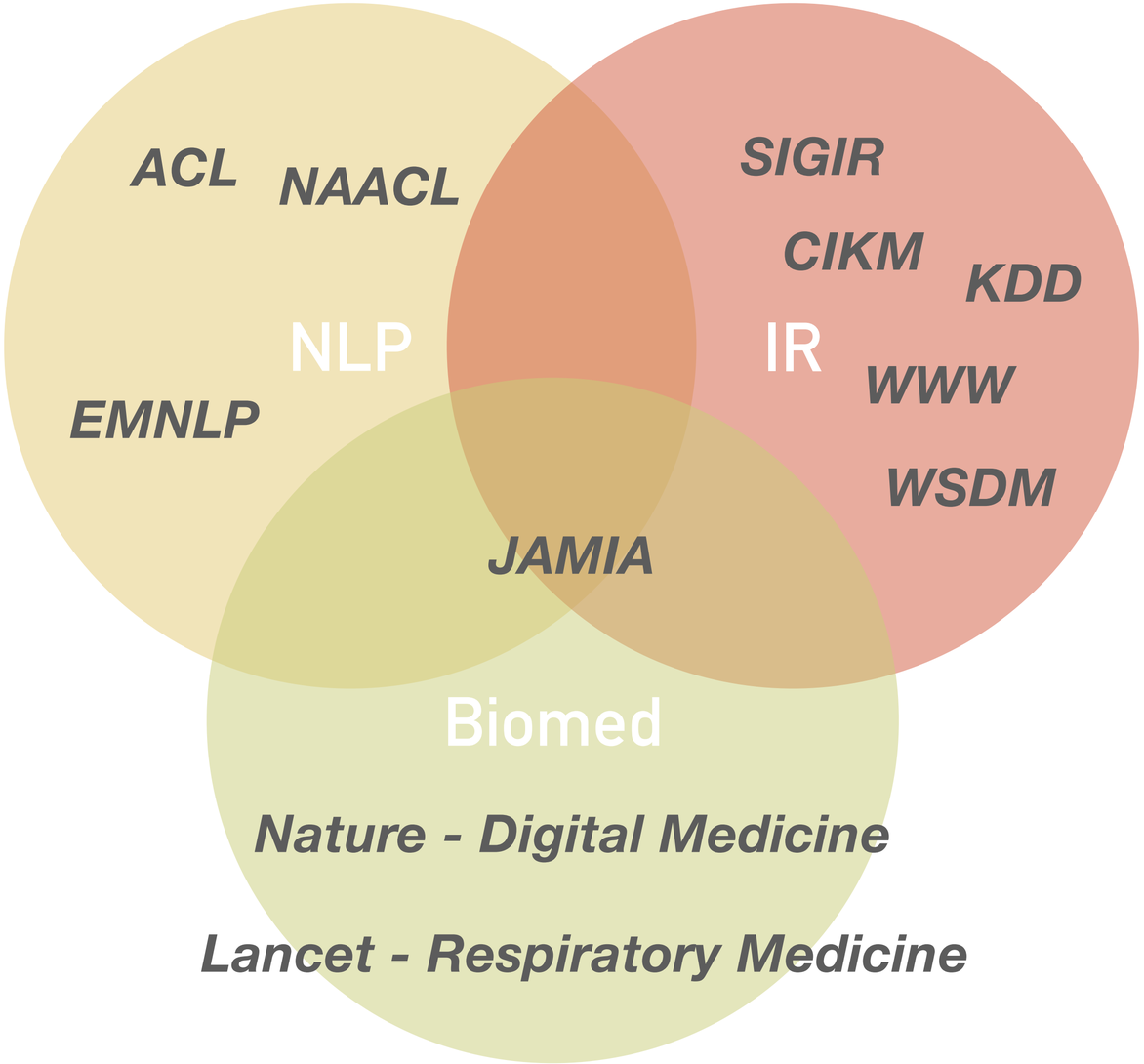
About

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



Interests

- **Dense Retrieval**
- **Explainable IR**
- **Uncertainty Aware IR**
- **Grounded Language Modeling**
- **Manifold Learning for Neural LMs**
- **Conditional Text Generation**
- **Clinical Decision Support**



Team

CARSTEN EICKHOFF
PROFESSOR

LAURA MERCURIO MD
ASSISTANT PROFESSOR

ALI BAHRAINIAN
POSTDOC

AUGUSTO GARCIA-AGUNDEZ
POSTDOC

GEORGE ZERVEAS
PHD STUDENT

JACK MERULLO
PHD STUDENT

MICHAL GOLOVANEVSKY
PHD STUDENT

RUOCHEN ZHANG
PHD STUDENT

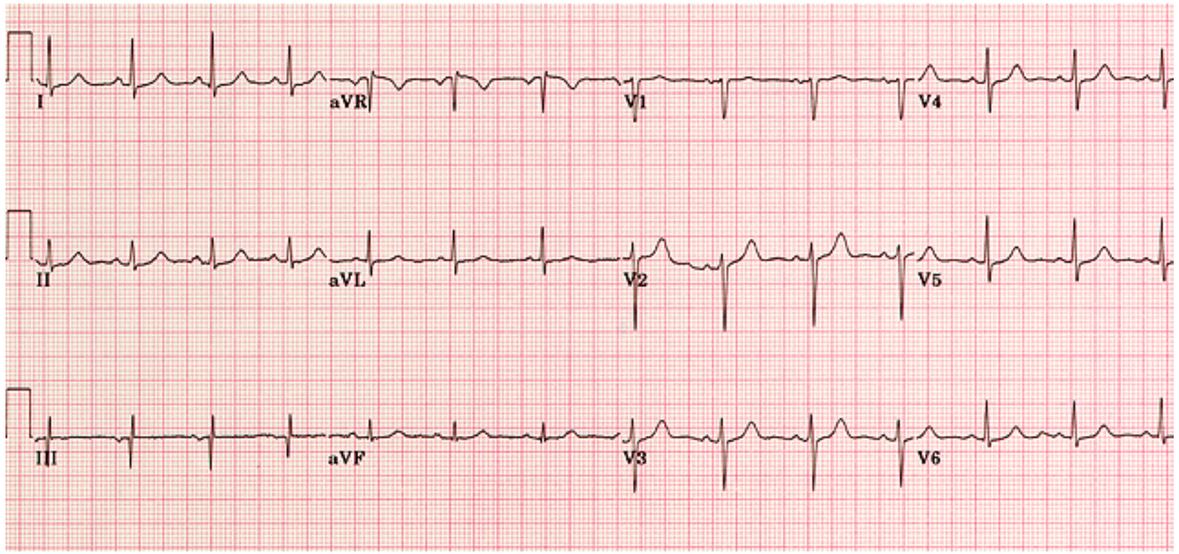
AMINA ABDULLAHI
PHD STUDENT

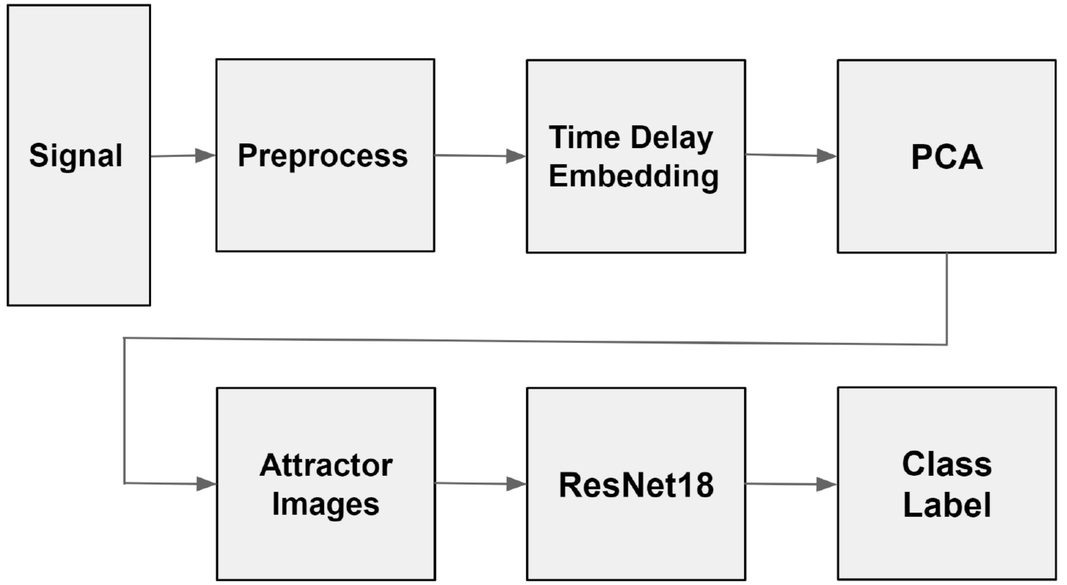
CATHERINE CHEN
PHD STUDENT

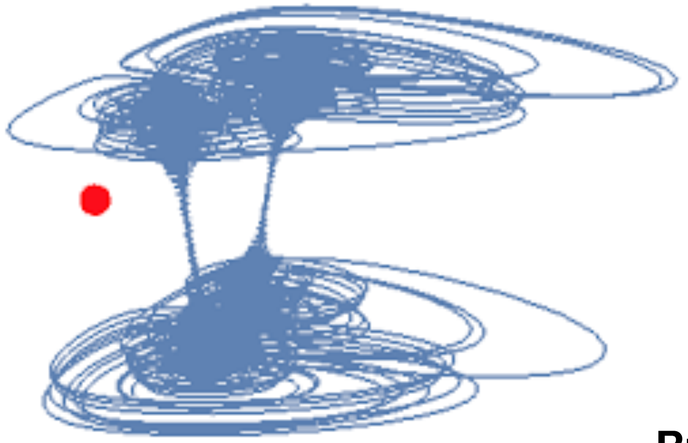
WASIWASI MGONZO
PHD STUDENT

WILLIAM RUDMAN
PHD STUDENT

Cardiac Arrhythmia Localization via ECG Attractor Imaging



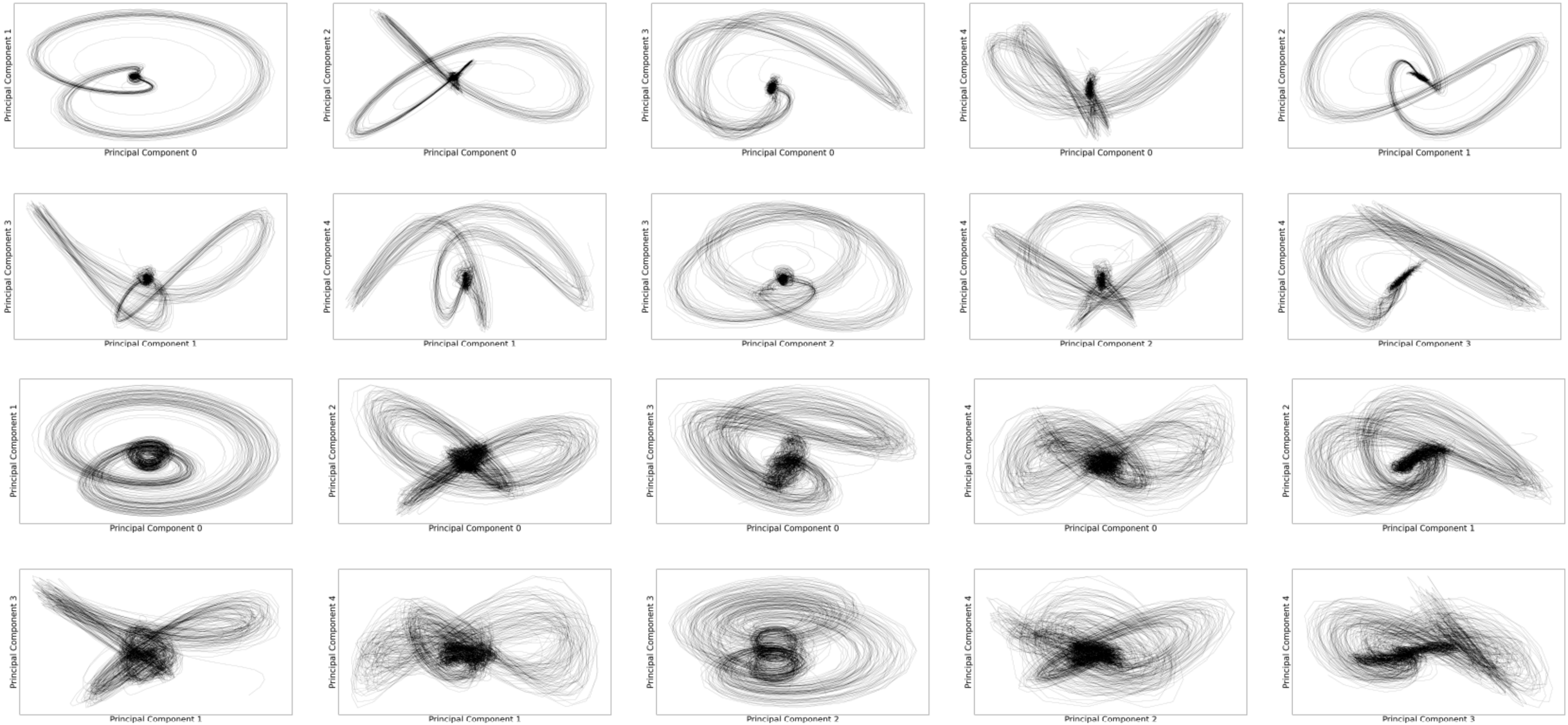




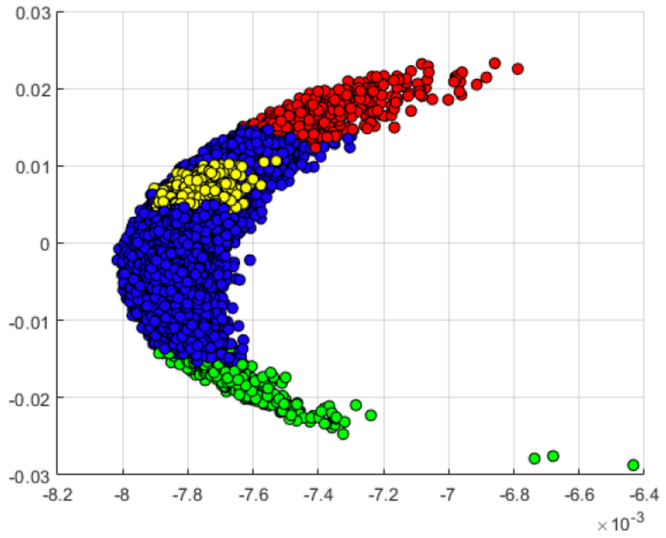
	Normal	AFib	Other	Noise	Avg. Score
Zabihi et. al. [33]	90.49	79.43	75.64	61.11	81.85
Datta et. al. [9]	91.00	79.00	77.00	–	–
Hong et. al. [15]	92.04	86.92	80.68	81.56	85.30
Spectrogram	76.78	43.08	44.71	54.55	54.78
Signal	94.84	96.77	91.93	90.41	93.49
Naïve Attractor	93.93	85.71	94.40	94.87	92.23
01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

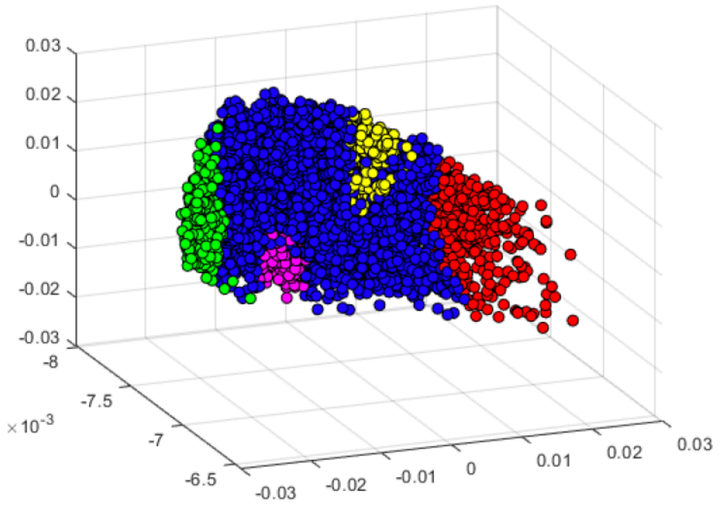
Rudman et al. CinC 2022: ACQuA: Arrhythmia Classification with Quasi-Attractors

Some Examples



Health Outcome Disparities on the EHR





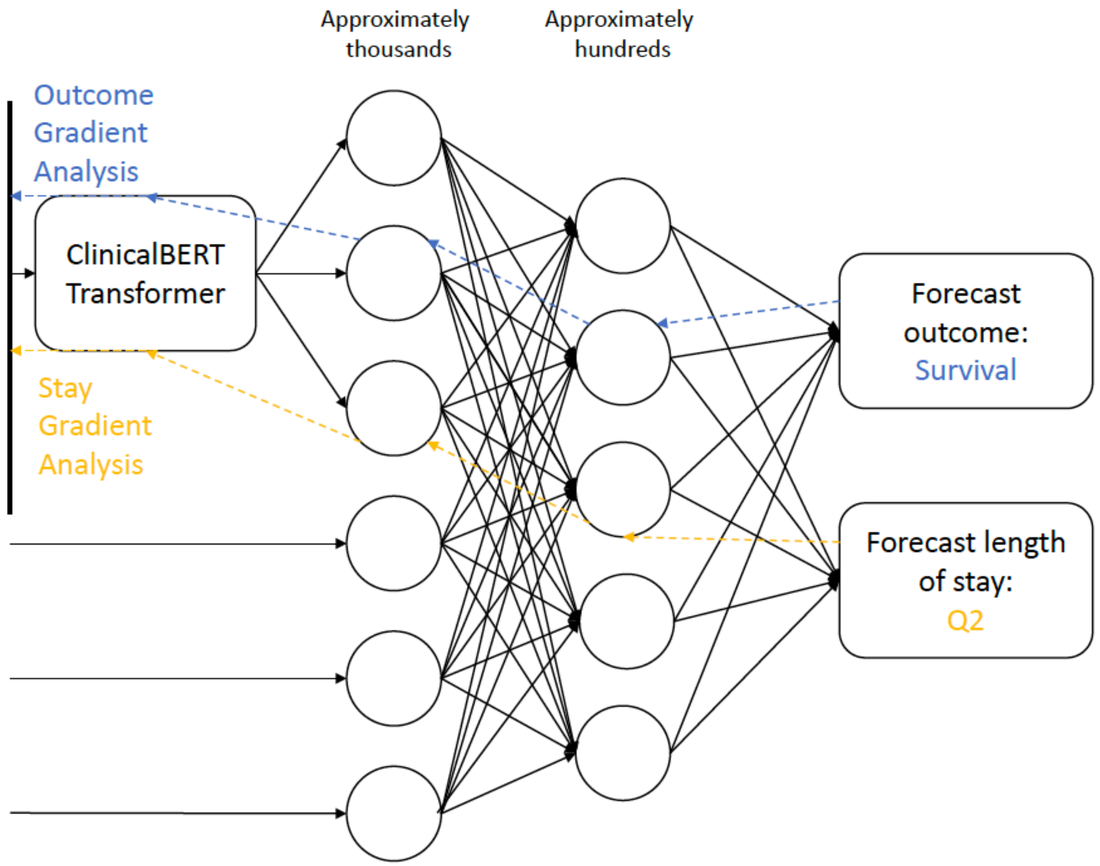
- Self Attention Contextual Embeddings

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

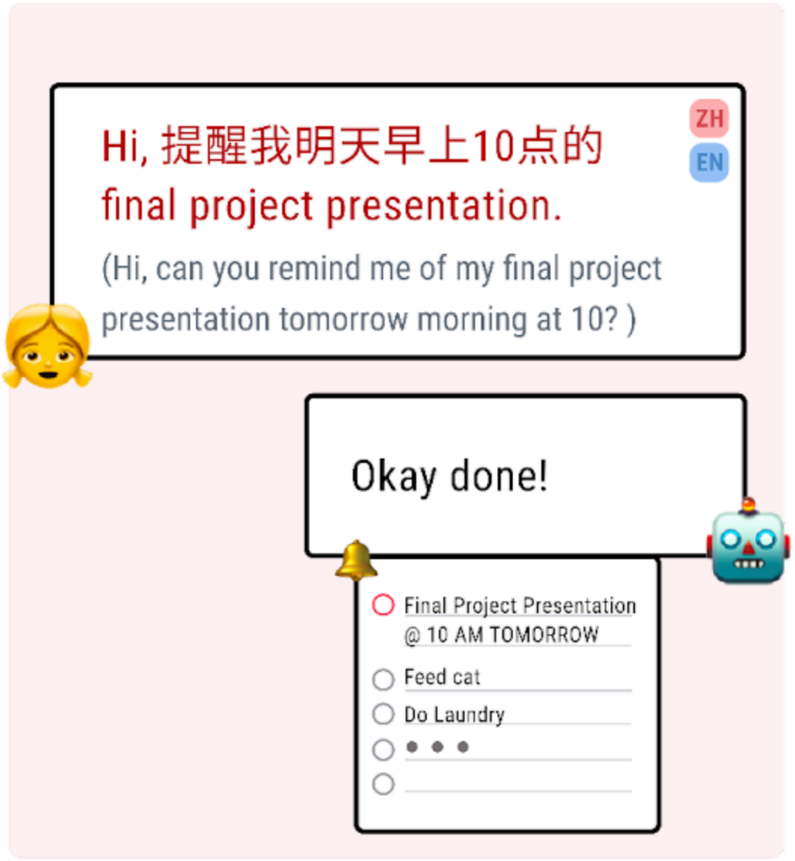
…

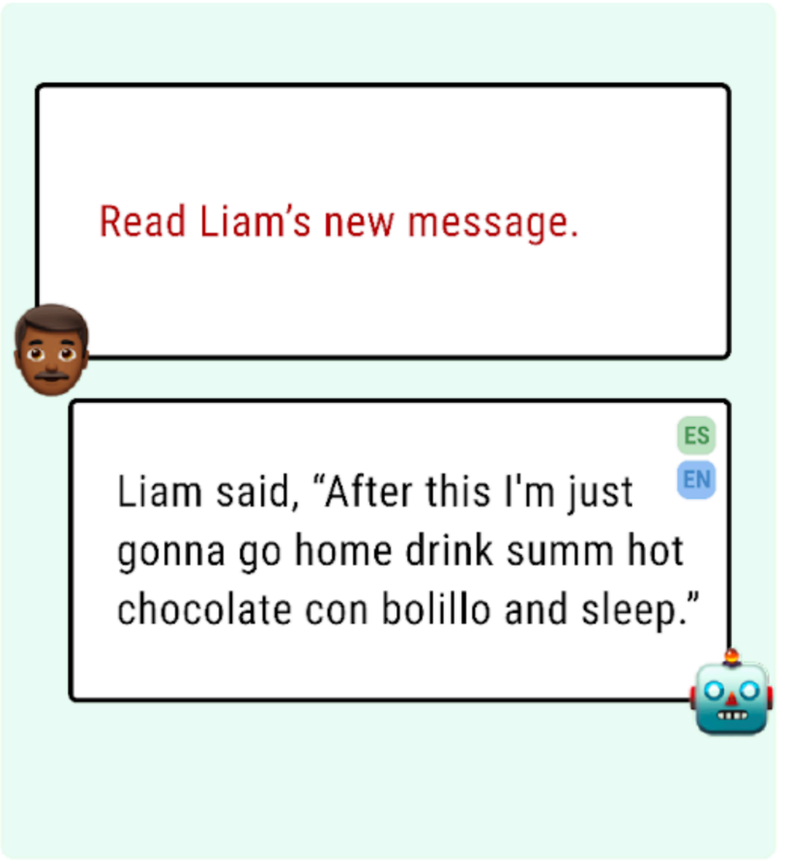
- Comorbidity n yes/no



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

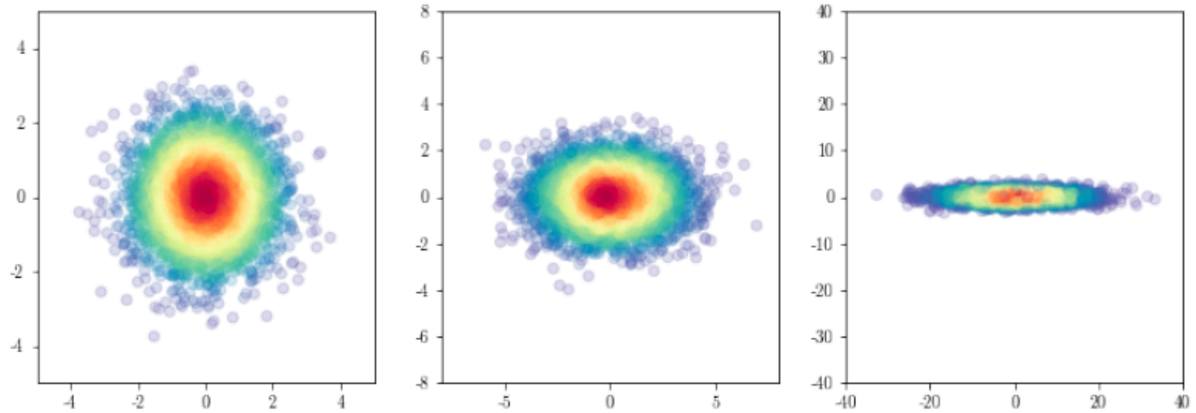
Language Generation in a Code-switched World

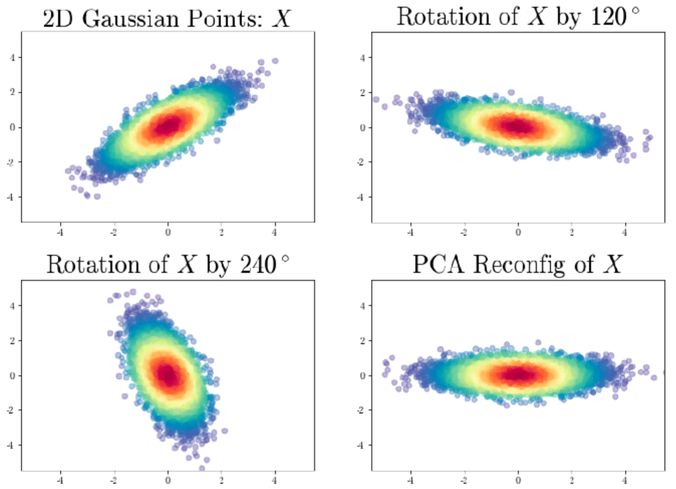




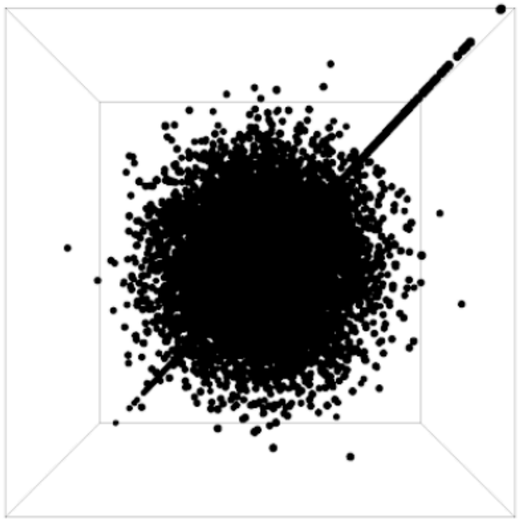
Zhang et al. *under review*: CroCoSum A Benchmark Dataset for Cross-Lingual Code-Switched Summarization

Topological Embedding Analysis





Test	IsoScore	AvgRandCosSim	Partition	ID Score	VarEx
1. Mean Agnostic	✓	✗	✗	✓	✓
2. Scalar Covariance	✓	✗	✗	✓	✓
3. Maximum Variance	✓	✗	✓	✗	✗
4. Rotation Invariance	✓	✓	✗	✓	✓
5. Dimensions Used	✓	✗	✗	✗	✗
6. Global Stability	✓	✗	✓	✓	✗



Rudman et al. ACL 2022: IsoScore: Measuring the Uniformity of Embedding Space Utilization

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

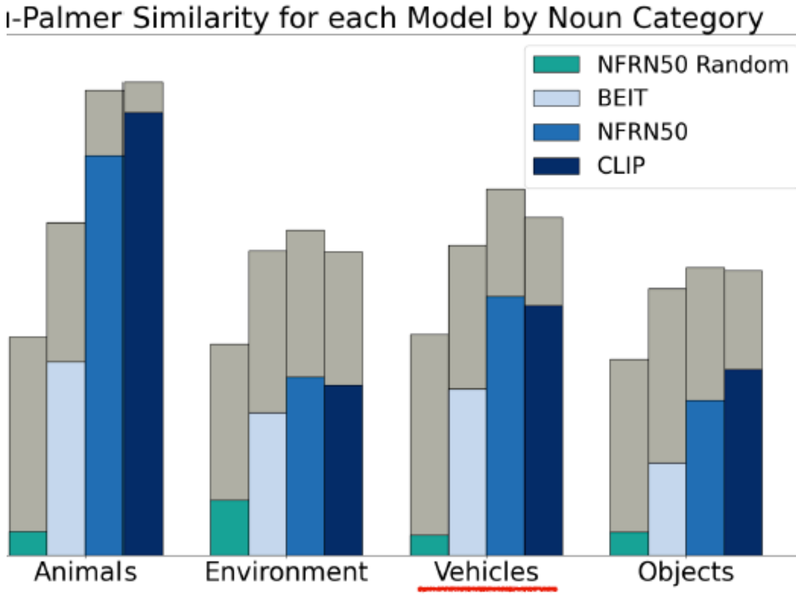
Visual Question Answering

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

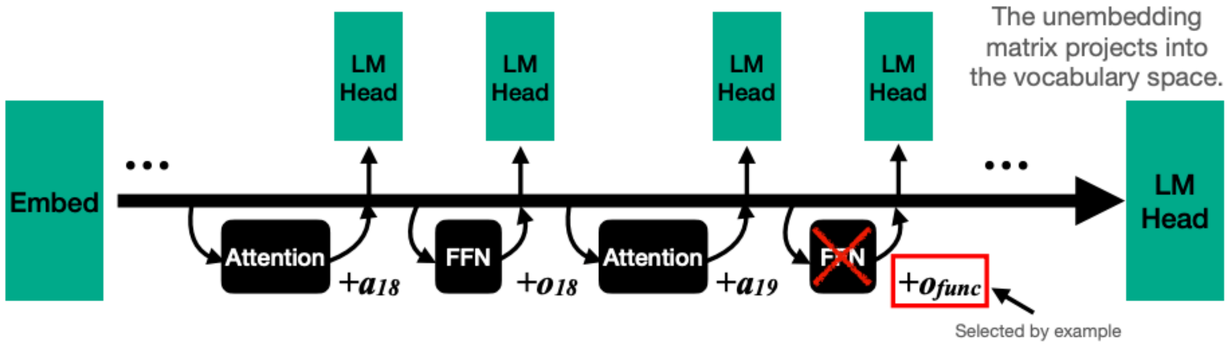


Merullo et al. ICLR 2023: Linearly Mapping from Image to Text Space

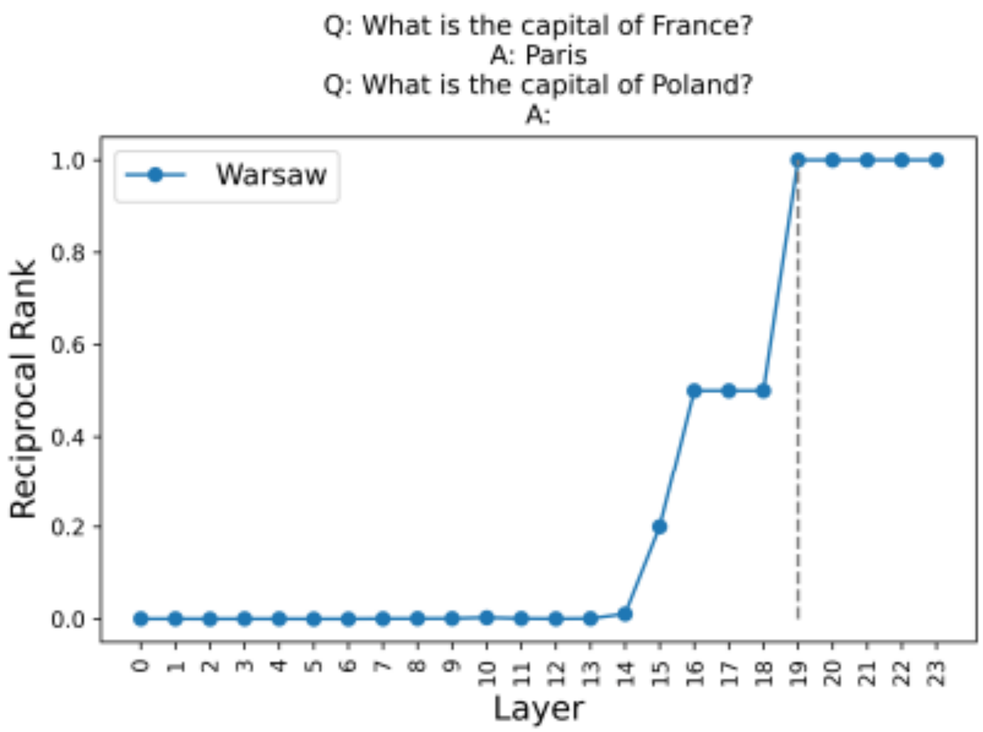
Grounding Language Models in Physics

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

LLM Vector Arithmetics

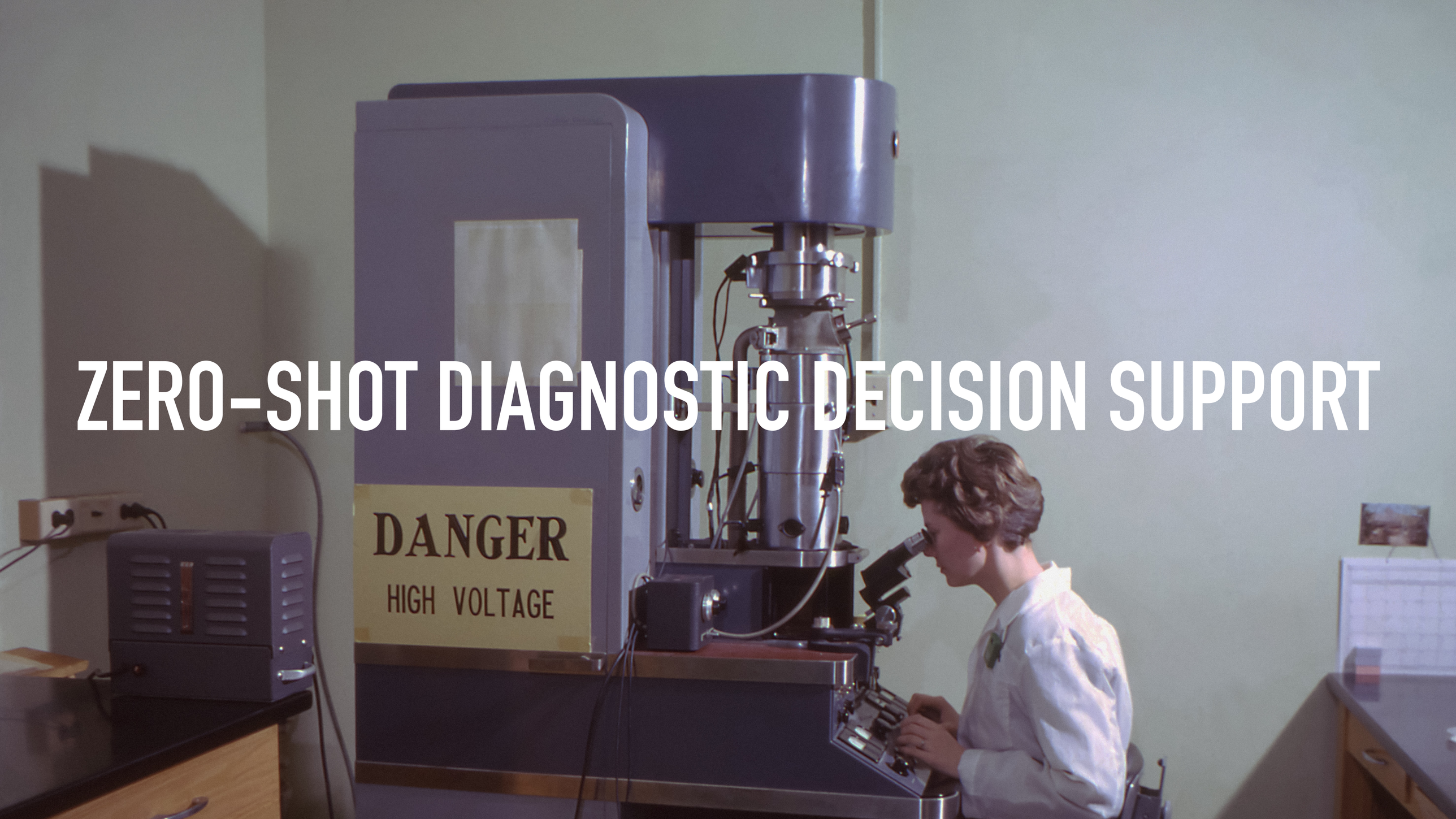


- ***We find simple vector arithmetic in Transformers***
- ***3 Models***
  - ***GPT2 (124M)***
  - ***GPT-J (6B)***
  - ***Bloom (176B)***

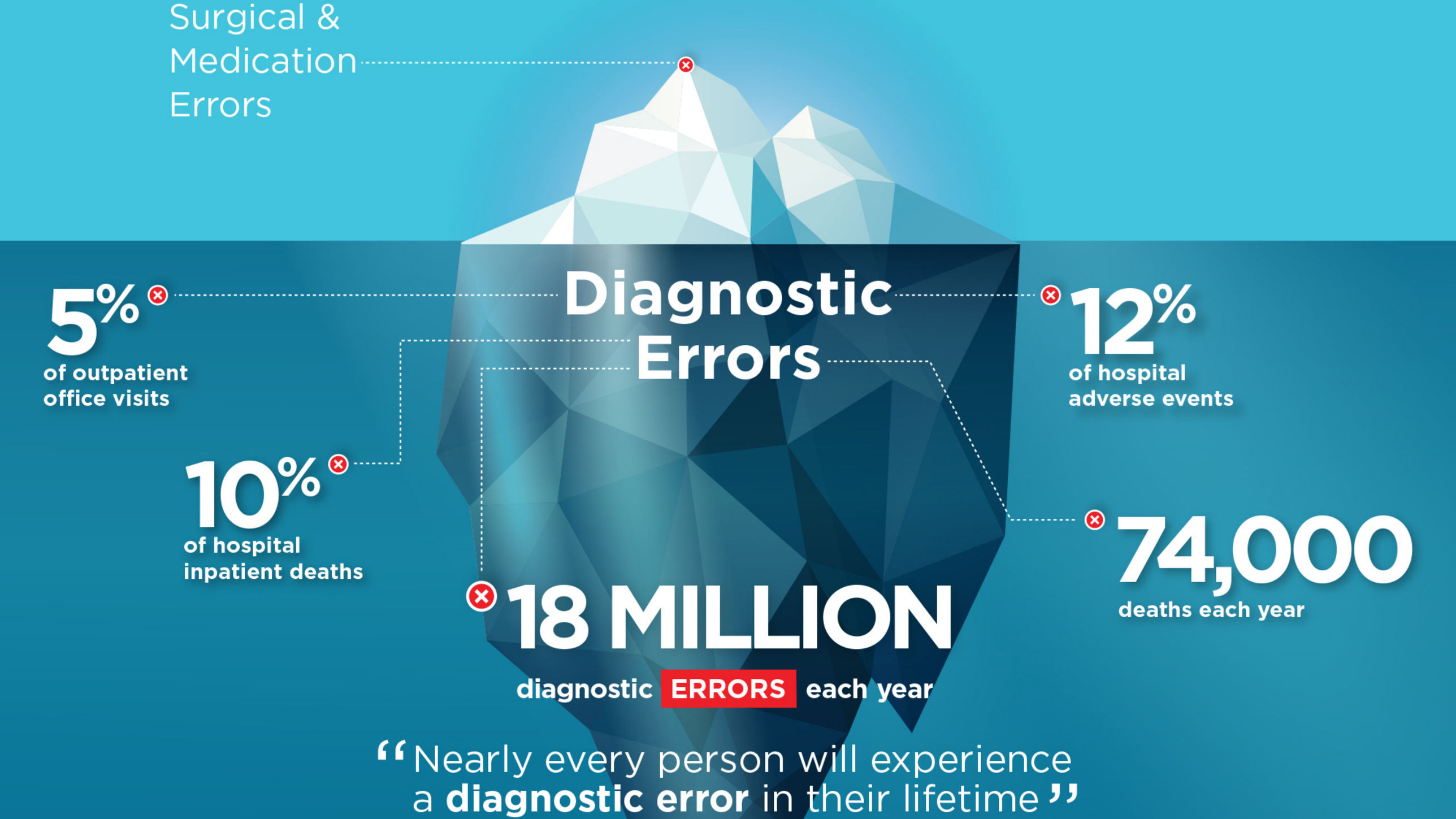


Layer	Top Token	
0	(	
1	A	
2	A	
3	A	
4	A	
5	A	
6	No	
7	C	
8	A	
9	A	
10	A	
11	A	
12	Unknown	
13	C	
14	St	
15	Poland	A
16	Poland	A
17	Poland	A
18	Poland	A
19	Warsaw	B
20	Warsaw	C
21	Warsaw	C
22	Warsaw	C
23	Warsaw	C

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

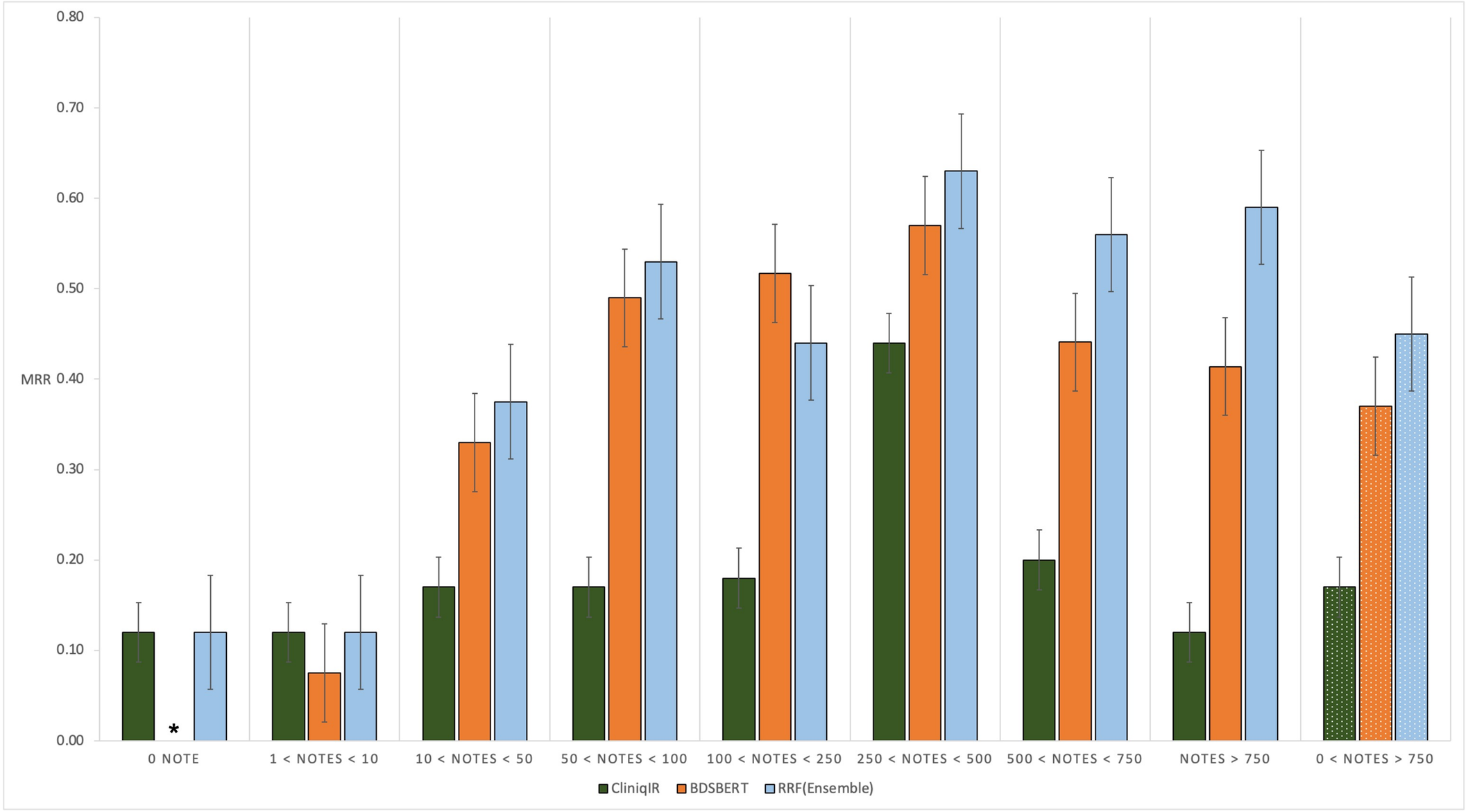
- *Golovanevsky et al. 2022*: Alzheimers (92.28%)
- Delahanty et al. 2018: Sepsis (97%)
- *Gulshan et al. 2016*: Diabetic Retinopathy (99.1%)
- *Rudman et al. 2022:* Cardiac Arrhythmias (99.27%)
- …

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]
- …

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)

Clinical LLMs

Week (05/15/23)- Can Large Language Models Perform Complex Diagnoses?

- We tested these models on the DC3 dataset
- It contains 30 difficult to diagnose case challenges

Models	Task I (Open-ended)	Task II (Multiple-choice)	Task III (Multiple-choice)	Total
ChatGPT	8	13	11	16
GPT4	8	**14**	**13**	17
BARD	**10**	**14**	**13**	**18**


HEALTH-NLP.ORG

Why is ChatGPT so good?

- Attention
- Lots of (diverse) training data
- Masked Language Modeling
- Tight Guardrails