***Time** Matters*: Examine Temporal Effects on Biomedical Language Models

Weisi Liu*, Zhe He^ and Xiaolei Huang*

* University of Memphis, TN, USA
^ Florida State University, Tallahassee, FL, USA

Time root in model development & deployment

Historical Data

Data temporal shift

Future Data

Questions about the flu?
Yes!

What about the COVID-19?

Sorry, this is beyond my knowledge.

Does the model still perform well? If not, why?

Current Study

PMC-LLaMA

Clinical T5

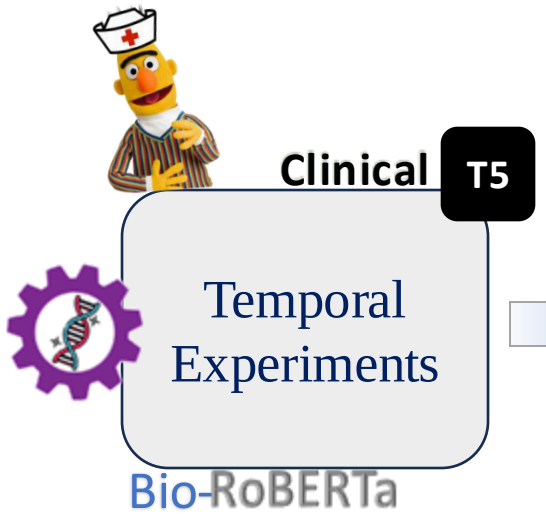
Bio-RoBERTa

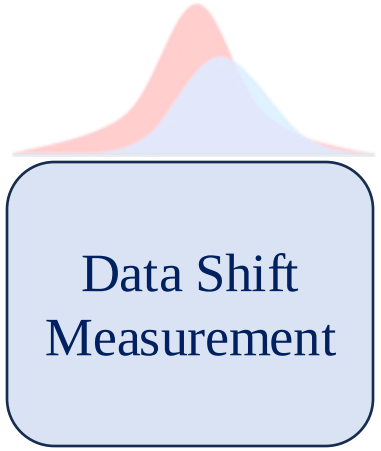
Time?

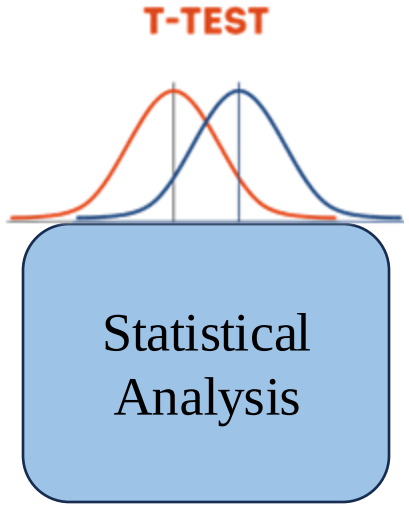
Rapid development of the LLMs

But, very **limited** study about model <u>temporal effects</u> in the biomedical domain…

Overview







Phenotype Inference

Information Extraction

Question Answering

Datasets

- MIMIC:

  2014-2019

Clinical Notes for Patient.

Predict ICD-10 codes → **'I10'** (Hypertension)

'E119' (Type 2 diabetes mellitus)

- BioNER:

  2009-2013

Addition of neutralizing anti **- TNF - alpha** antibodies drastically reduced **p24** antigen release and prevented **CD4** + cell depletion associated with infection.

B-Protein, I-Protein, E-Protein

- BioASQ:

  2014-2023

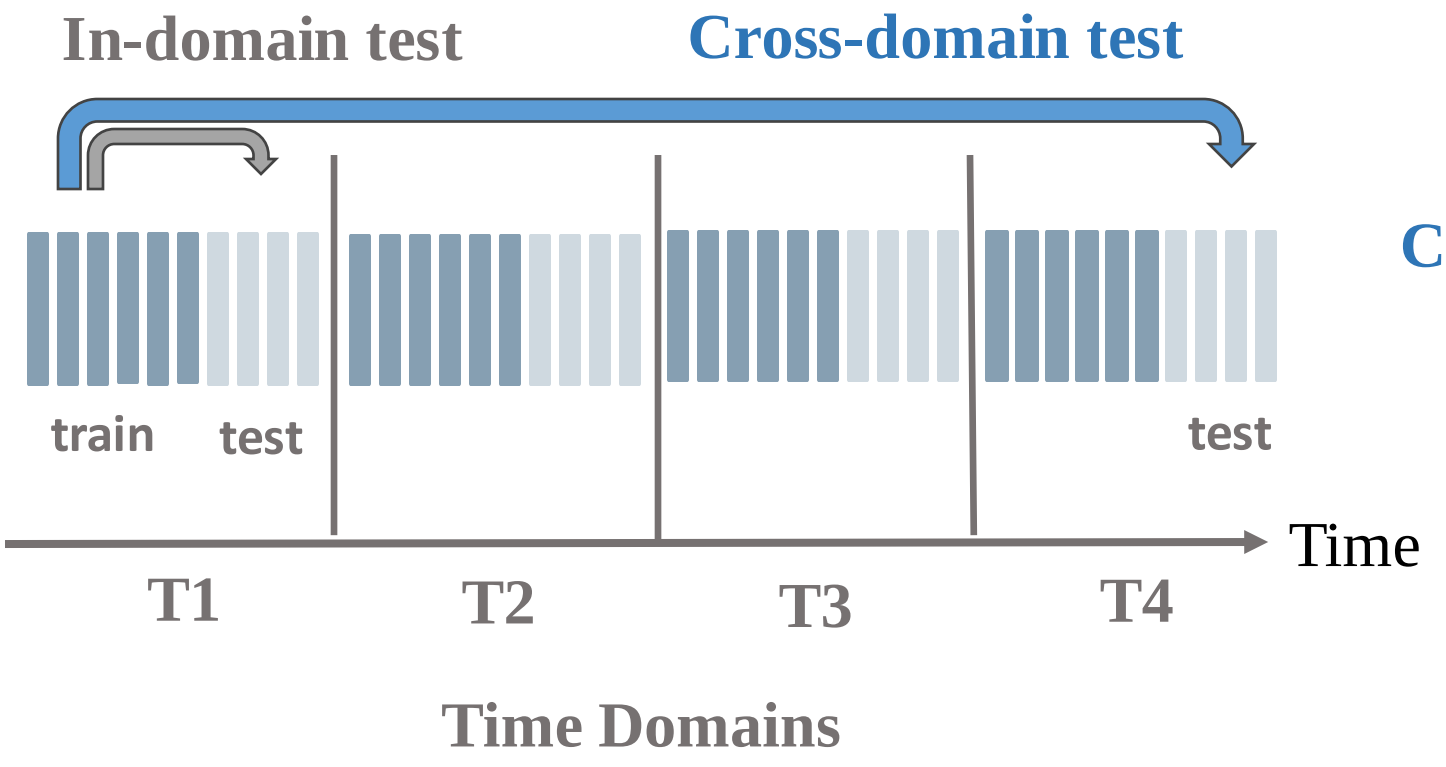
Question: Where in the cell do we find the protein Cep135?

+

Snippets form scientific articles related to the question.

Answer: "centrosome"

Experiments:
Model Temporal Performance Variation



Performance Change:

Cross-domain test performance
minus
In-domain test performance

Data Shift Measurement

Word level Metric

Jaccard Similarity

TF-IDF Cosine Similarity

Semantic level Metrics

Use **Encoder Models** to obtain semantic representation of time domains, and measure the shift (e.g. Cosine similarity, Euclidean distance).

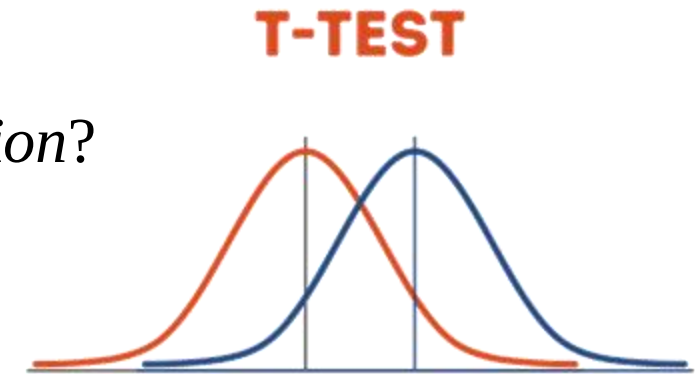
Universal Sentence Encoder [1]
SBERT [2]
BioLORD [3]
MedCPT [4]

[1] Cer D, et al. Universal sentence encoder.
[2] Reimers N, et al. Sentence-BERT: Sentence embeddings using siamese BERT-networks.
[3] Remy F, et al. BioLORD-2023: Semantic textual representations fusing large language models and clinical knowledge graph insights.
[4] Jin Q, et al. MedCPT: Contrastive pre-trained transformers with large-scale PubMed search logs for zero-shot biomedical information retrieval.

Statistical Analysis

Two-tailed T-test:

- Does the model really have *performance degradation*?
- Is there really a *data shift*?



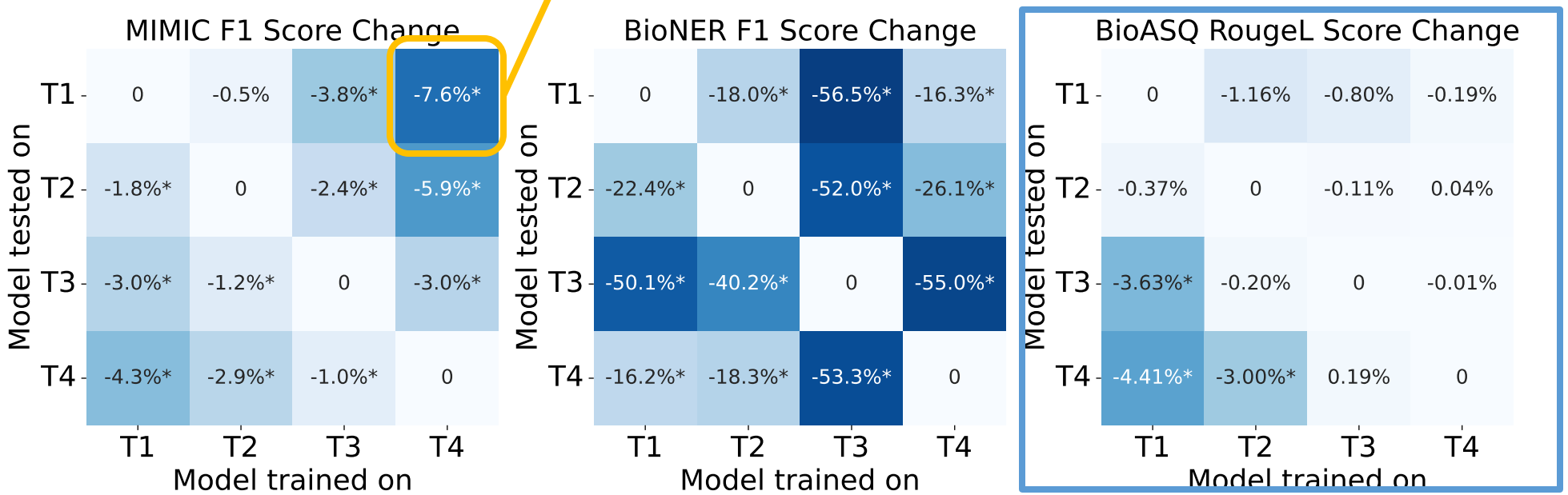
Pearson Correlation Coefficient:

- What's the correlation between *data shift* and model *performance variations*?

Results & Discussions

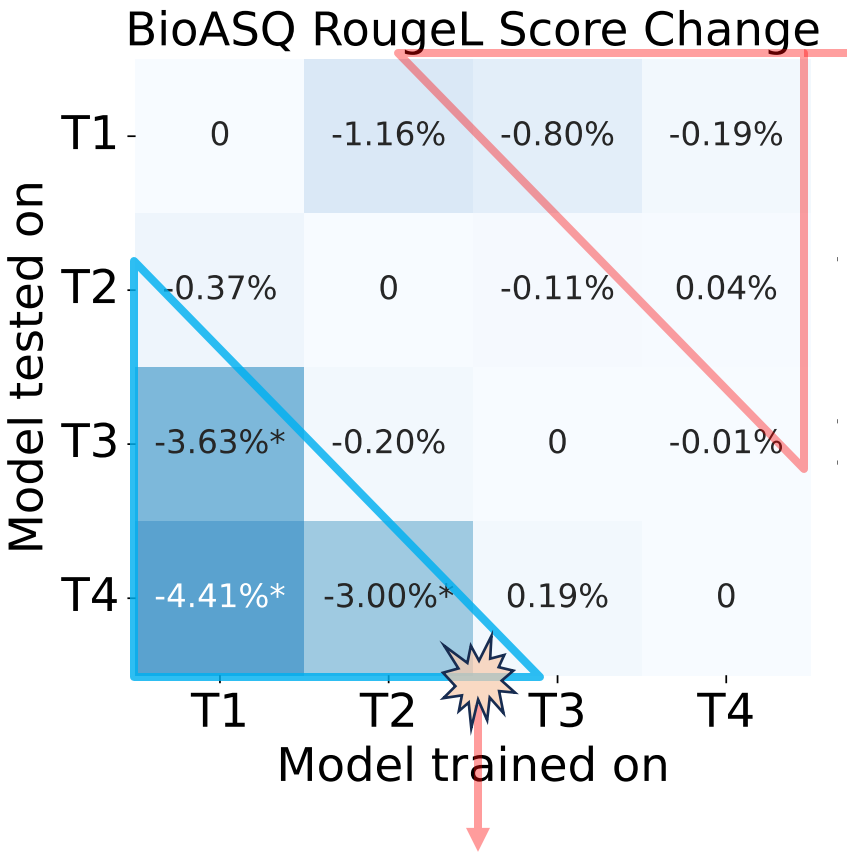
Q1. Does the performance of biomedical language models change over time?

Performance change of model trained on T4 and tested on T1



Performance change heatmaps. The star (*) indicates the performance change is significant.

Specific temporal event can impact model performance across time!



COVID-19 at this point

Does not show statistically significant degradation

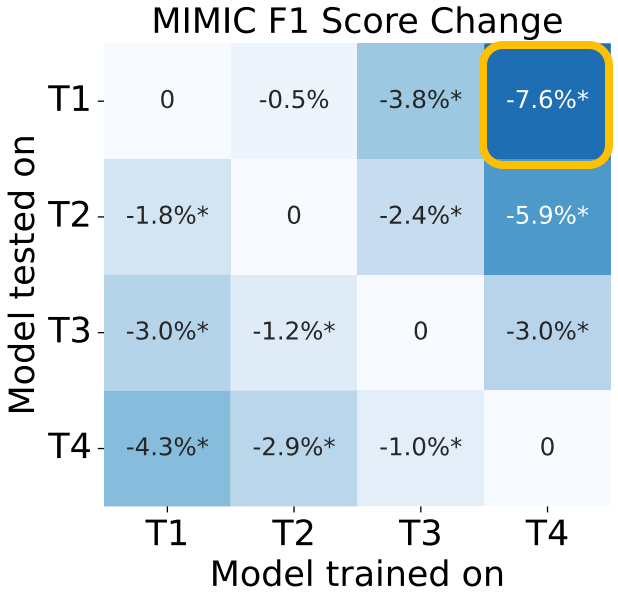
specific temporal event?

Verification Experiment:
A small COVID-19 related test dataset derived from BioASQ test split (BioASQ-COVID)

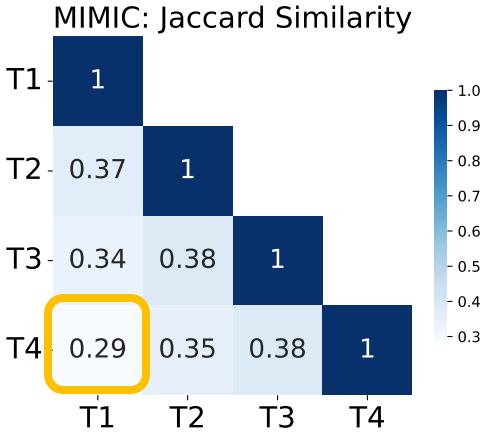
Model	T1	T2	T3	T4
RougeL(%)	37.6	35.2	**38.8**	**41.9**

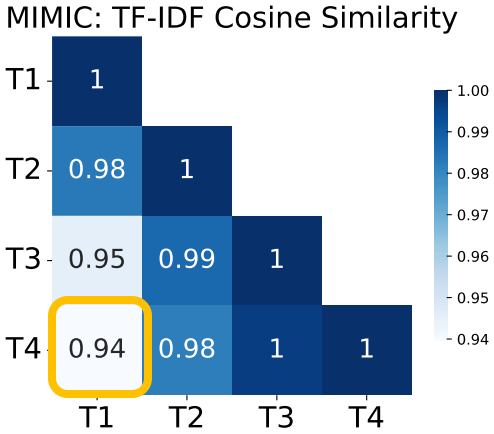
Performance (RougeL) of models trained on different time domain and test on BioASQ-COVID

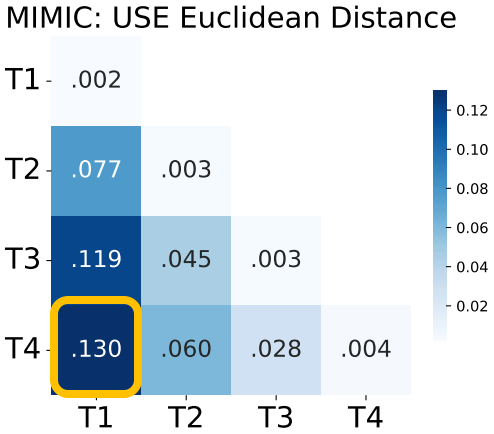
*Q2: Does the **performance change** statistically correlate with the **data shift**?*



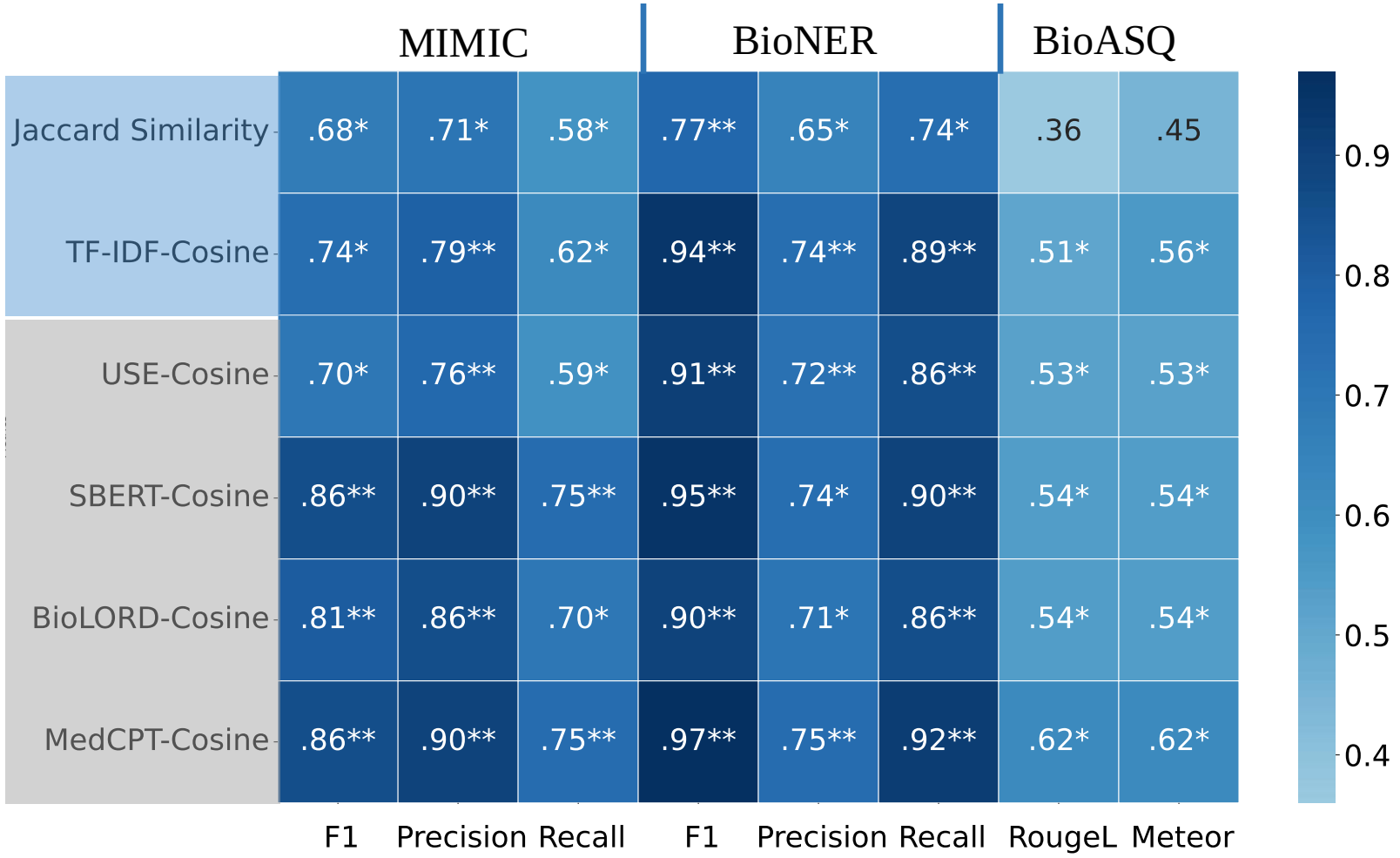
Performance change heatmap.







Heatmaps of data shift measurement across domain pairs in MIMIC



Yes!

But…

The heatmap of correlation coefficient between the model performance changes and data shift measurements Over the three dataset. * indicates the p-value is less than 0.05, and ** indicates the p-value is less than 0.001

Q3: Do All Data Shift Measurements Tell Us the Same Story?

Q3: Do All Data Shift Measurements Tell Us the Same Story?

	MIMIC			BioNER			BioASQ	
	F1	Precision	Recall	F1	Precision	Recall	RougeL	Meteor
Jaccard Similarity	.68*	.71*	.58*	.77**	.65*	.74*	.36	.45
TF-IDF-Cosine	.74*	.79**	.62*	.94**	.74**	.89**	.51*	.56*
USE-Cosine	.70*	.76**	.59*	.91**	.72**	.86**	.53*	.53*
SBERT-Cosine	.86**	.90**	.75**	.95**	.74*	.90**	.54*	.54*
BioLORD-Cosine	.81**	.86**	.70*	.90**	.71*	.86**	.54*	.54*
MedCPT-Cosine	.86**	.90**	.75**	.97**	.75**	.92**	.62*	.62*

0.9
0.8
0.7
0.6
0.5
0.4

Word level metrics

Semantic level metrics

General domain encoders
v.s.
Biomedical encoders

Takeaways

1. Biomedical models are time-sensitive
   - Timely knowledge updates or model adaptation is needed

2. Data shift evaluation could be an indicator
   - Signals when model adaptation is necessary

3. Different metrics provide varying perspectives on data drift
   - Choose suitable metrics based on task nature

Scan the QR code for more details in our paper at:
https://arxiv.org/pdf/2407.17638

Email: wliu9@memphis.edu

Acknowledgement:

NIH NSF AHRQ Agency for Healthcare Research and Quality