Synthetic Data and Large Language Model Applications

COERE Methods

John D. Osborne, PhD
Associate Professor
University of Alabama at Birmingham
April 10th, 2025

What is Synthetic Data?

"*information that is artificially generated rather than produced by real-world events*"

–Wikipedia

(In contrast to processed or unprocessed data generated from the actual event)

Why Synthetic Data?

Cheaper

Faster

Privacy

Imaginary People with Imaginary PHI!

Bias

Another, less explored tool in the bias mitigation toolbox

Possible Disadvantages

Realism

Set-up Costs

Wrong labels

“Model Collapse”

Types and Examples of Synthetic Data

Templated

- Generated by a deterministic algorithm
- Often used to fill known gaps in the data
- Combinations and permutations of existing data may be used
- Ex) Synthea, Faker library

Generative

- Stochastic process
- Created through Generative Models
  - Language Models
    - Transformers, RNNS…
  - Diffusion Models
    - Stable Diffusion, DALL-E
  - Reinforcement learning
- Ex) GPT-4

Two Problems that Synthetic Data Can Help With

Surrogate Substitution for De-identification

- Osborne, John D., Tobias O'Leary, Akhil Nadimpalli, and Richard E. Kennedy. "Bratsynthetic: Text de-identification using a markov chain replacement strategy for surrogate personal identifying information." *arXiv preprint arXiv:2210.16125* (2022).

Biomedical Entity Linking or Normalization

- Kuleen Sasse, Shinjitha Vadlakonda, Richard E. Kennedy and John D. Osborne. “Disease Entity Recognition and Normalization is Improved with Large Language Model Derived Synthetic Normalized Mentions”. *arXiv preprint arXiv:2410.07951* (2024).

Problem 1:

What Surrogates Should Replace PHI in De-identified Text?

What is a Surrogate?

Original Text:
MRN: 123456 OSBORNE, JOHN 10/11/2024

PHI and PII

- PHI: Personal Healthcare Information
- PII: Personal Identifying Information

Possible Surrogate Values

- MRN: ###### XXX, XXX ##/##/####
- MRN: ###### XXXXXXX 11/11/2024
- MRN: 999999 DOE, JOHN 01/01/2001
- MRN: {MRN} {PATIENT_NAME} {DATE}

What is De-identification?

Safe Harbor

- Full de-identification method
- Removes all PHI including names, MRNs, ages over 89, dates (except year), locations smaller than a State, etc…
- All information that “could be used alone or in combination with other information to identify an individual”

Limited Data Set

- Removes many of the same elements
- Keeps dates including birth year and smaller geographic locations to facilitate research
- Information is still considered to contain PHI
- Data Use Agreement is required

Both methods require surrogate generation

How to do De-Identification?

- Find PHI
  - Named Entity Recognition Tasks
  - Encoder models are SotA
- Replace PHI
  - Options traditionally are ENTITY_NAME or consistent substitution of “realistic text”

Highlights

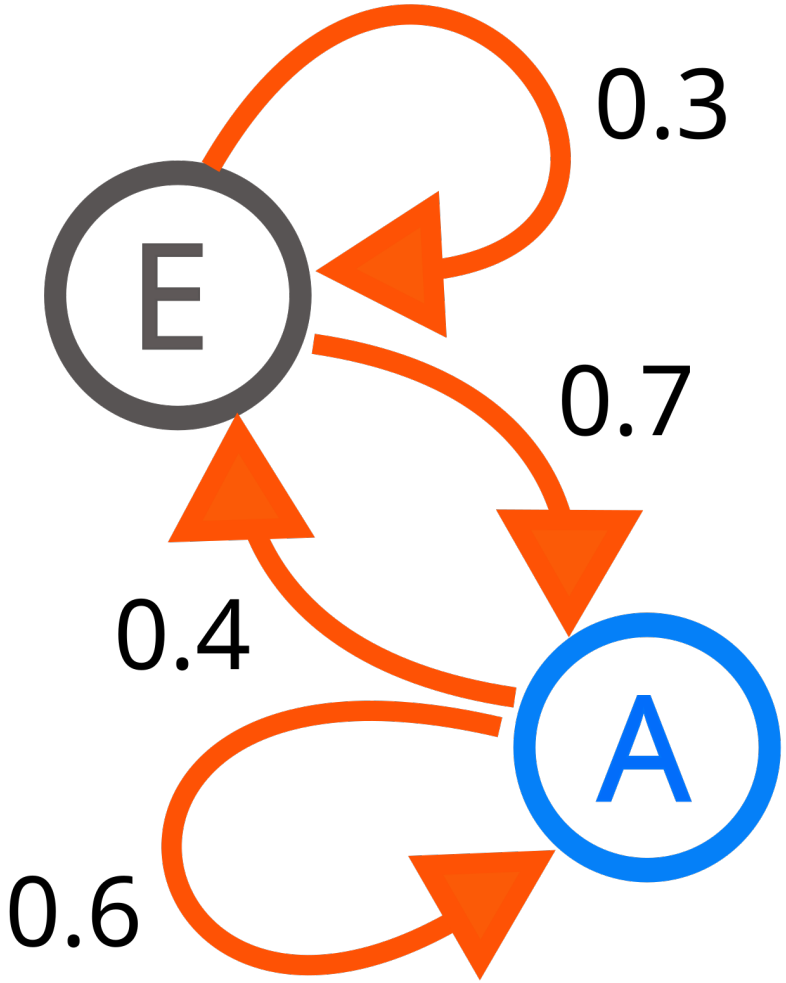
- De-identification is a key step in making EHR data accessible for further research.
- Machine learning and hybrid approaches are predominant de-identification methods.
- The majority of the approaches were trained and evaluated on public corpora.
- Current state-of-the-art systems provide binary token F1-scores of over 98%.
- Limited diverse annotated corpora and domain adaptation methods pose challenges.

Kovačević, Aleksandar, Bojana Bašaragin, Nikola Milošević, and Goran Nenadić. "De-identification of clinical free text using natural language processing: A systematic review of current approaches." *Artificial Intelligence in Medicine* (2024): 102845.

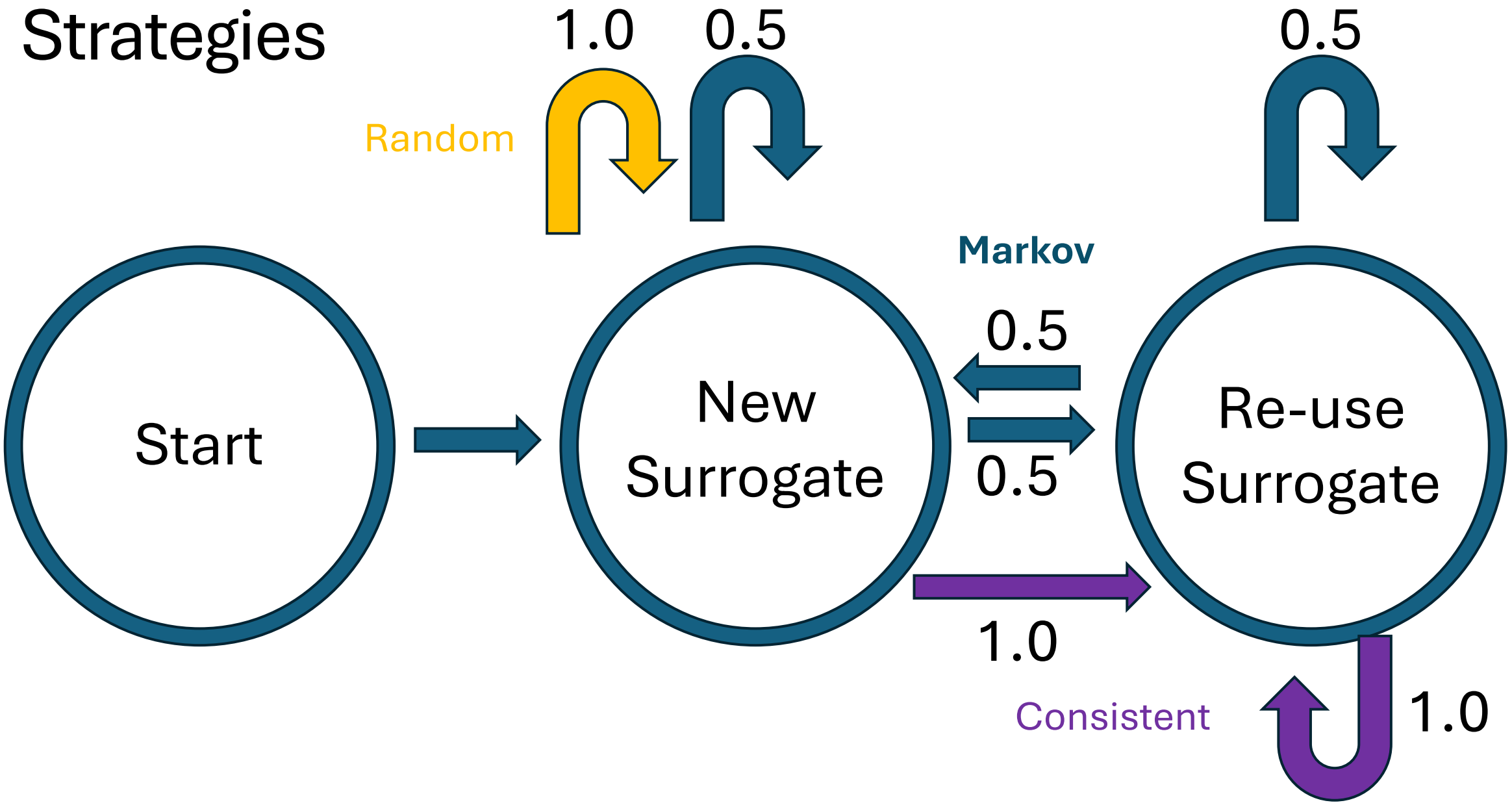
What is a Markov Process?

Definition

- A **Markov chain** or **Markov process** is a stochastic process describing a sequence of possible events in which the probability of each event depends only on the state attained in the previous event.
  - Wikipedia



Substitution Strategies



	Substitution Strategy			
Substitution	**None**	**Consistent**	**Random**	**Markov**
None (Original Name)	Sandy	Sandy	Sandy	Sandy
1st Substitution	ENTITY_NAME	Sara	Kim	Sara
2nd Substitution	ENTITY_NAME	Sara	Nisha	Sara
3rd Substitution	ENTITY_NAME	Sara	Cathy	Ann
4th Substitution	ENTITY_NAME	Sara	Maria	Maria
5th Substitution	ENTITY_NAME	Sara	Hannah	Maria
6th Substitution	ENTITY_NAME	Sara	Lin	Maria
False Negative !!	**Sandy**	**Sandy**	**Sandy**	**Sandy**

Maximum Alternative Surrogate Repetition (MASR)

	Substitution Strategy			
Substitution	**None**	**Consistent**	**Random**	**Markov**
None (Original Name)	Sandy	Sandy	Sandy	Sandy
1st Substitution	ENTITY_NAME	Sara	Kim	Sara
2nd Substitution	ENTITY_NAME	Sara	**Nisha**	Sara
3rd Substitution	ENTITY_NAME	Sara	Cathy	Ann
4th Substitution	ENTITY_NAME	Sara	Maria	**Maria**
5th Substitution	ENTITY_NAME	Sara	Hannah	**Maria**
6th Substitution	ENTITY_NAME	Sara	Lin	**Maria**
False Negative !!	Sandy	Sandy	Sandy	Sandy
MASR	**0**	**0**	**1**	**3**

Method

De-identification & Surrogate Substitution

- De-identification of UAB notes with BERT-based Named Entity Recognition software
  - MIMIC notes already de-identified…
- Surrogate substitution with BRATsynthetic
  - https://github.com/uabnlp/BRATsynthetic
  - Date offsetting
  - Multiple surrogate substitution strategies
  - Uses faker library
    - https://github.com/joke2k/faker

Data Sets

- UAB Corpus
  - All available EHR notes for 165 patients from 2014 to 2021
- MIMIC Corpus
  - Critical Care Notes
  - SemEval 2015 Task 14 subset

Supplementary Table 3. UAB and MIMIC Critical Entity Distribution

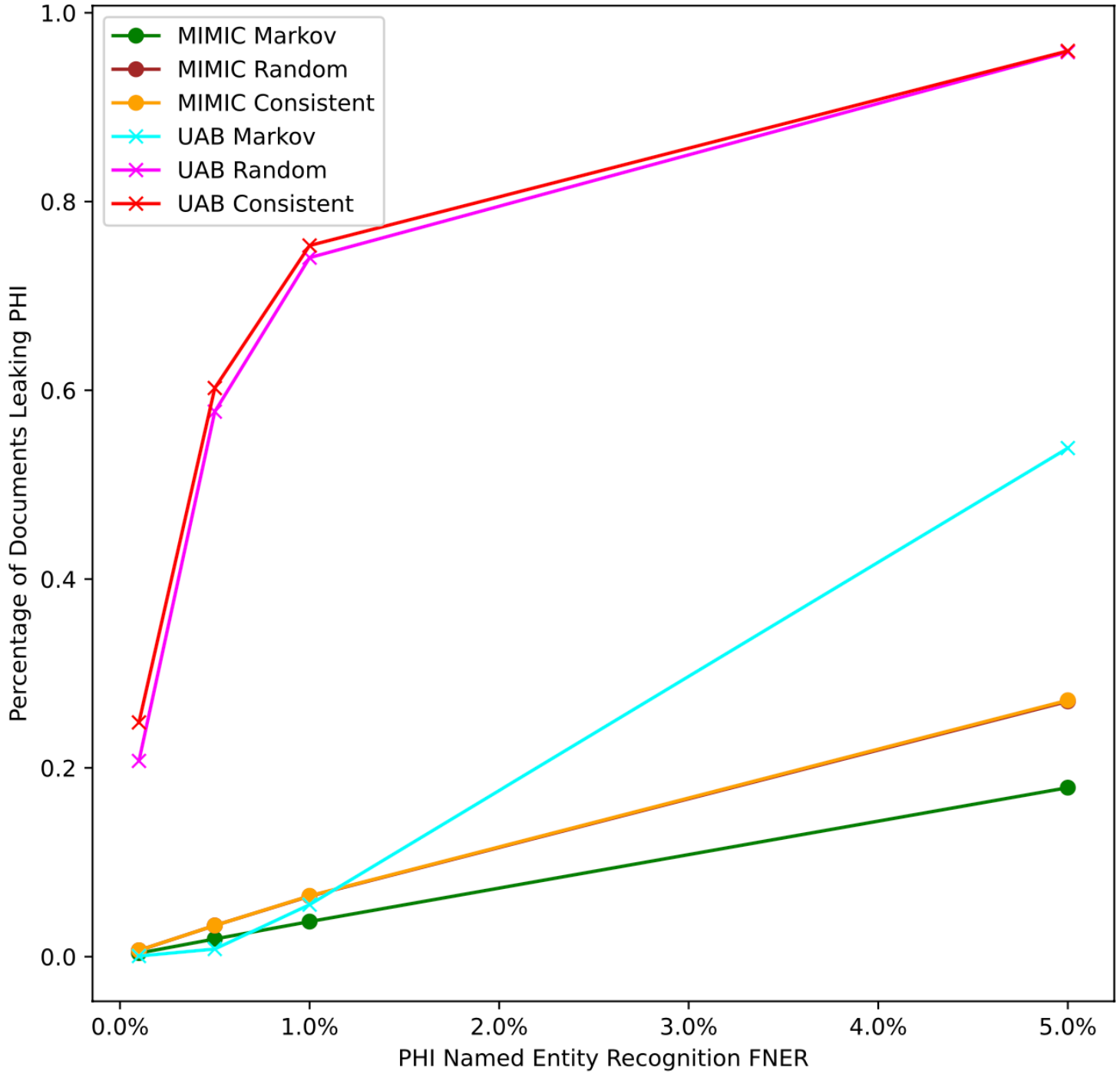
Document Statistics				Patient Statistics	
	UAB Total	UAB Discharge	MIMIC Discharge		UAB Total
Critical Entities Mean	388.5	355.6	6.8	Critical Entities Mean	8123.0
Critical Entities Median	224	199	5	Critical Entities Median	985
Critical Entities Range	2 – 2545	10-2414	2-76	Critical Entities Range	7 - 321945

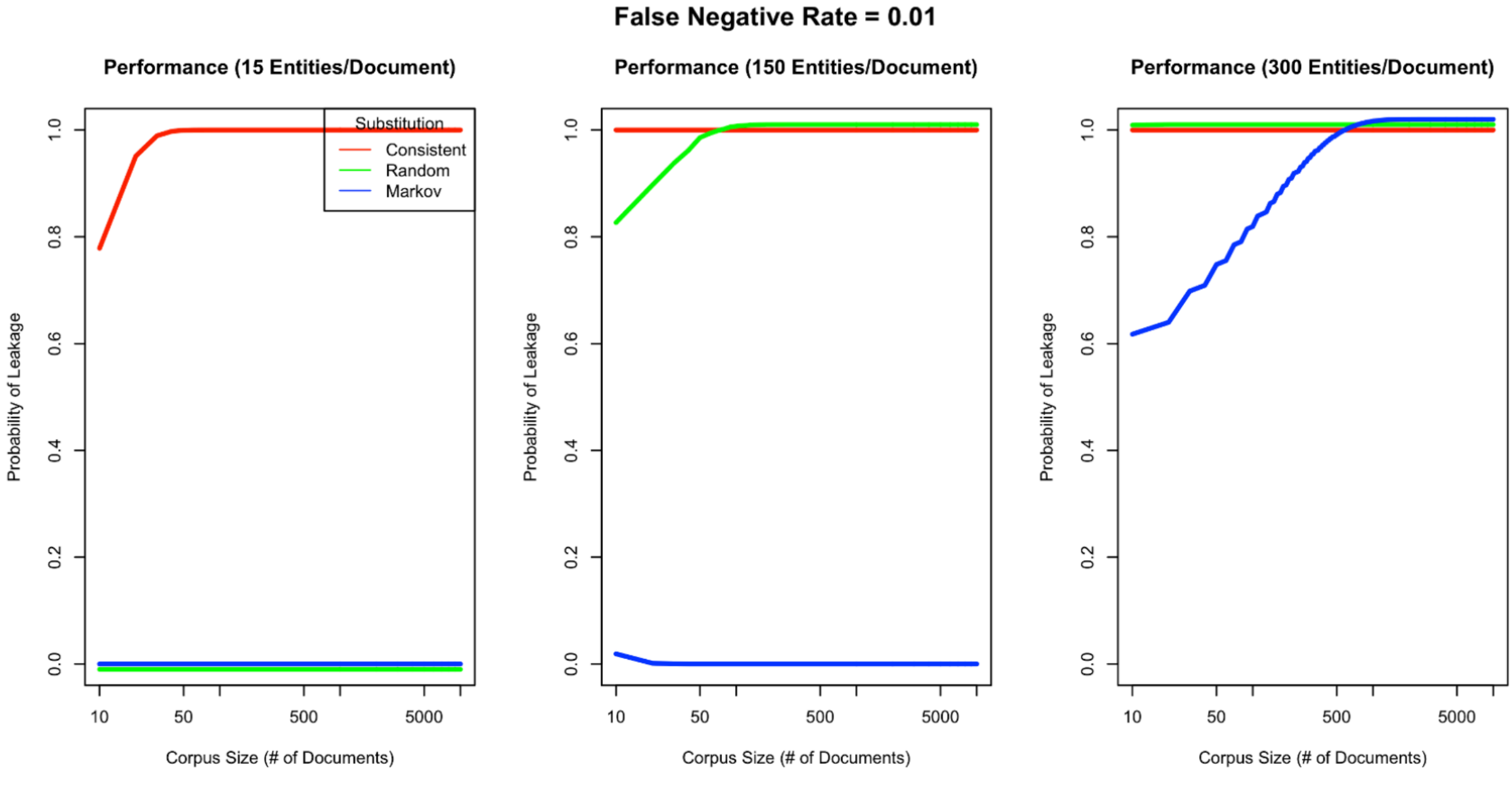
Method: Estimating PHI Leakage

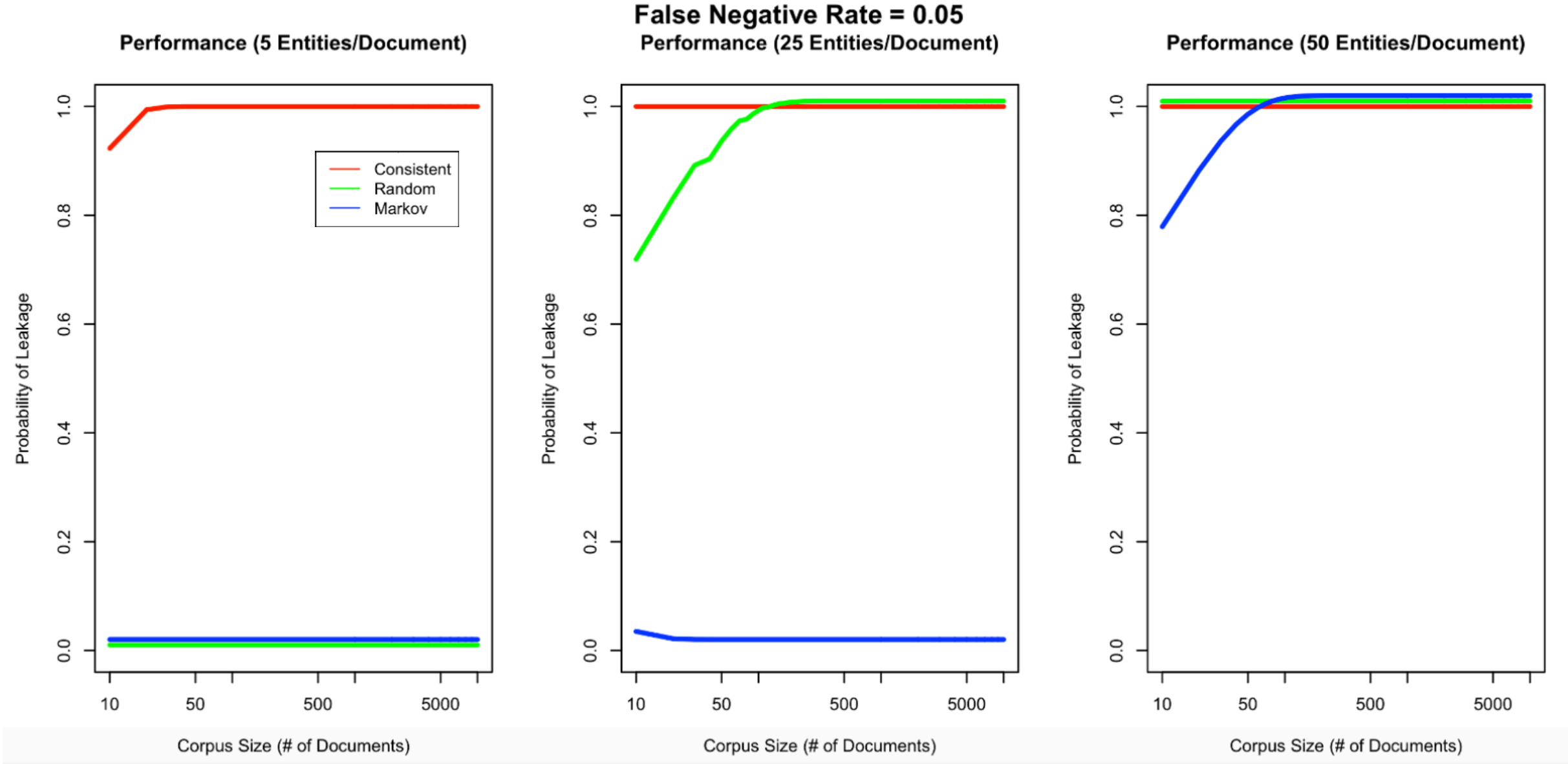
- Leakage was said to occur when the expected number of real patient entities (FN errors) in the corpus was greater than a set threshold for each of the substitution methods
  - 0 for consistent
- For Random and Markov substitution, the threshold was set as the expected number of fake entities based on a pool of 1,000 fake entities from which to randomly select using the transition probabilities described above

Real World Data

UAB vs MIMIC Corpus PHI Document Leakage Rates




False Negative Rate = 0.01
Performance (15 Entities/Document)
Performance (150 Entities/Document)
Performance (300 Entities/Document)
Substitution
Consistent
Random
Markov
Probability of Leakage
0.0
0.2
0.4
0.6
0.8
1.0
10
50
500
5000
Corpus Size (# of Documents)


False Negative Rate = 0.05
Performance (5 Entities/Documents)
Performance (25 Entities/Documents)
Performance (50 Entities/Documents)
Consistent
Random
Markov
Probability of Leakage
0.0
0.2
0.4
0.6
0.8
1.0
10
50
500
5000
Corpus Size (# of Documents)

But this inconsistent stuff will mess up NLP?!

	Substitution Strategy			
Substitution	**None**	**Consistent**	**Random**	**Markov**
None (Original Name)	Sandy	Sandy	Sandy	Sandy
1st Substitution	ENTITY_NAME	Sara	Kim	Sara
2nd Substitution	ENTITY_NAME	Sara	Nisha	Sara
3rd Substitution	ENTITY_NAME	Sara	Cathy	Ann
4th Substitution	ENTITY_NAME	Sara	Maria	Maria
5th Substitution	ENTITY_NAME	Sara	Hannah	Maria
6th Substitution	ENTITY_NAME	Sara	Lin	Maria
Similarity to Original	**Lowest**	**High**	**Low**	**Intermediate**

Method

- Evaluate synthetic PHI documents on downstream tasks
  - 3rd Party (TriNetX Genomic Pipeline)
  - Clinical Modifiers (Opioid Use Disorder Pipeline)
  - POS Tagging & Dependency Parse (MedSpacy Pipeline)
- Data Set
  - De-identified UAB Opioid Use Disorder Data Set
- Work in Progress
  - Re-run on updated data set
  - Inclusion of Simple substitution strategy

TriNetX Genomic Analysis

TriNetX Pipeline Methodology

1. Tokenization
2. Dictionary Lookup for finding mentions in the document
3. Disambiguation with linear SVMs considering the context
4. Rules for specific patterns
5. Result classification with linear SVMs

Data and Results

- 3067 UAB OUD Notes
- Only 31 assessed as having genomic related content
- 5 notes initially had differences, reduced to 4 after it was discovered one note was not run

Error Analysis

Document Type	Orig	Cons	Rand	Mark	Explanation
DWN - Psychiatry/Psychology	0	1	0	0	ER: UAB ER, -> ER: ACH ER
RAD - MR Breast Diagnostic Bil wo+w contrast	2	2	1	2	Estrogen receptor positive status [ER+] (True Positive)
DWN - Emergency Department	0	0	0	2	recently seen in ER on 4/4
DWN - Emergency Department	0	1	2	2	Pt was recently seen in ER on 02/08
DWN - Consult Notes	15	15	15*	16	?

- Changing the name of the ER from UAB ER to ACH ER creates false positive, but UAB ER to EBI ER or CMC ER does not!
- This pipeline is working with **out of domain data**, it has never seen UAB data before

NER and Clinical Modifiers for OUD

	None	Consistent	Random	Markov
NER	0.723	0.706	<u>0.731</u>	0.702
Subject	<u>0.936</u>	0.918	0.926	0.927
DocTime	0.926	0.925	<u>0.934</u>	0.925
Negation	0.969	0.975	<u>0.975</u>	0.972

- Random slightly outperforms the original text on NER and 2/3 of the clinical modifiers
- This pipeline is working with **in domain data**, it has seen UAB data before

Spacy: POS Tags and Dependency Parse

Minimal Change vs Original

	PUNCT	ADP	VERB	ADJ
Consist.	651079	131406	120639	194866
Random	651135	131420	120706	194442
Markov	651131	131439	120629	194706
Original	650552	131648	120690	195509

A Lot of Change vs Original

	ADV	nummod	ROOT	conj
Consist.	27633	250030	86173	14753
Random	30860	86180	159028	17607
Markov	30872	250339	8971	14743
Original	27615	248188	85642	87790

- Much more change than expected
- This pipeline is working with **out of domain data**, development didn't include UAB data or clinical data

Discussion

- NLP is impacted in some cases (TriNetX) or Spacy's POS Tagger (see below)
  - Impact appears to say more about model fragility than deficiences in surrogate generation

Type	Mention		
Consistent	36 Years	Ethanol level 07/17/19 08:48	Ampheta 07/17/19
Markov	41 Years	Ethanol level 07/14/19 23:59	Ampheta 07/12/19
Random	37 Years	Ethanol level 07/13/19 22:17	Ampheta 07/17/19
Simple	[AGE][AGE]	Ethanol level [DATE][TIME]	Ampheta [DATE]

Table A5. Part of Speech comparison (*Blue is PropN, Red is Noun, Green is ADJ, Orange is Verb and Purple is NUM*)

Future Direction include LLM as evaluation agent and exploration of use to mitigate bias

Problem 2:
How Can We Better Identify Rare Diseases and Conditions in Biomedical Text?

Relevant Problems in the Disease Space

Disease Entity Recognition

- There may be multiple ways of referring to a disease or condition of interest
  - Delirium
  - Risky behavior
- “Patient believes it is the year 1900”
- shoots up with dirty needles because, “don’t care”

Disease Entity Normalization

- Map condition of interest to a vocabulary, ex) SNOMED-CT, Human Phenotype Ontology
- Precision Medicine has this problem with rare disease diagnosis, first step is often mapping to PATO

Hypothesis: Synthetic text mentions generated from large language models can assist with Disease Entity Recognition (DER) and Disease Entity Normalization (DEN)

Why disease?

	Disease CUIs			
	With	**Without**	**CUI Average**	**Total**
Synonyms	217252	102129	2.84916	909967
Definitions	53432	265949	0.21725	69417

Table A1 Distribution of Synonyms and Definitions for CUIs in the UMLS 2019AB Disease Semantic Group. There are 319381 Disease Group CUIs.

How and Why?

How?

- Identity a mention of a biomedical entity of interest, disease/condition, drug/chemical, gene, etc.. in text
  - **Named Entity Recognition**
- Normalize the mention, but assigning or linking it to a vocabulary or ontology
  - **Entity Normalization**

Why?

- Many applications
  - KG Construction
  - Drug repurposing
  - Phenotyping
  - Surveillance
- Normalizing is a critical step
  - Tylenol/ Acetaminophen
  - AD/Alzheimer's
  - NOTCH1 also known as hN1; AOS5; TAN1; AOVD1

Notable Tools for Identification

Named Entity Recognition

- Encoder type models (BERT variants) are widely used
  - BERT
  - RoBERTa
  - deBERTa
- LLMs are improving
- Hybrid models with knowledge injection

Entity Normalization

- Harder Problem
- Popular Tools
  - QuickUMLS
  - MetaMap
  - CTAKES
  - SAPBert
  - KrissBERT
- Leveraging knowledge resources / ontologies is critical given huge number of classes and minimal training data

QuickUMLS

- Uses ”CPMerge” to find approximate string similarity between dictionary entries and text
- Starts with character trigrams and progressively merges

```
from quickumls.spacy_component import SpacyQuickUMLS

# common English pipeline
nlp = spacy.load('en_core_web_sm')

quickumls_component = SpacyQuickUMLS(nlp, 'PATH_TO_QUICKUMLS_DATA')
nlp.add_pipe(quickumls_component)

doc = nlp('Pt c/o shortness of breath, chest pain, nausea, vomiting, diarrrhea')

for ent in doc.ents:
    print('Entity text : {}'.format(ent.text))
    print('Label (UMLS CUI) : {}'.format(ent.label_))
    print('Similarity : {}'.format(ent._.similarity))
    print('Semtypes : {}'.format(ent._.semtypes))
```

Method		*Prec*	*Rec*	*F-1*	*ms/doc*
MetaMap		0.49*	0.48*	0.48*	19,295*
cTAKES		**0.71**	0.53*	0.61	3,852*
QuickUMLS	$\alpha = 0.6$	0.50*	**0.75**	0.60	1,594*
	$\alpha = 0.7$	0.60*	0.66*	**0.63**	680*
	$\alpha = 0.8$	0.63*	0.60*	0.61	332*
	$\alpha = 0.9$	0.64*	0.56*	0.60	193*
	$\alpha = 1.0$	0.67*	0.54*	0.60	**143**

Table 1: Results for the i2b2 dataset. cTAKES outperforms QuickUMLS in precision, but QuickUMLS has better recall. QuickUMLS is 2.5 to 135 times faster than MetaMap or cTAKES. * indicates statistically significant differences from best value (Welch's t-test, $p < 0.05$).

- MetaMap is a legacy tool from 2001, not really for clinical text
- cTAKES is focused on clinical text
- QuickUMLS is faster, but performance has a lot of room for improvement

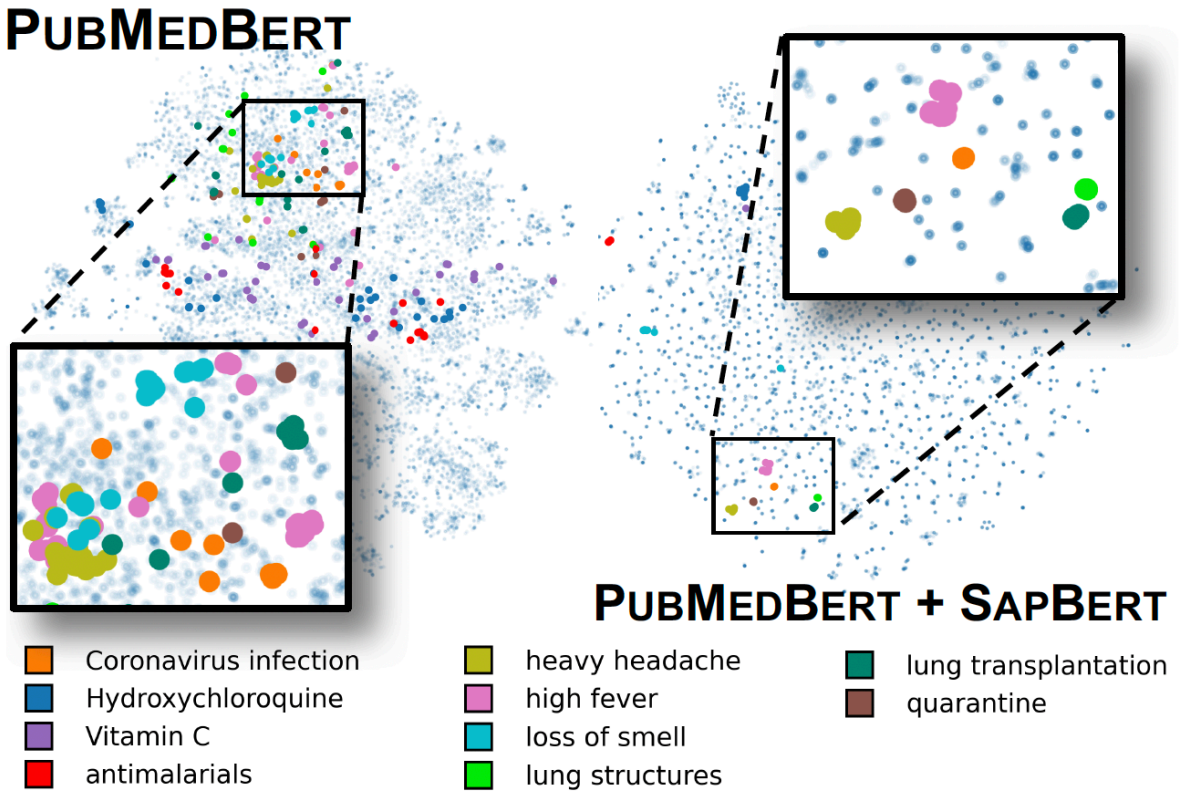


Figure 1: The t-SNE (Maaten and Hinton, 2008) visualisation of UMLS entities under PUBMEDBERT (BERT pretrained on PubMed papers) & PUBMEDBERT+SAPBERT (PUBMEDBERT further pretrained on UMLS synonyms). The biomedical names of different concepts are hard to separate in the heterogeneous embedding space (left). After the self-alignment pretraining, the same concept's entity names are drawn closer to form compact clusters (right).

- SapBERT
  - Self alignment pretraining
- Top system ~2021
- Leverages UMLS Data, but no LLMs
  - Mention, UMLS Concept Name Tuples
- Doesn't really normalize the way I would like… counts getting synonym as correct!

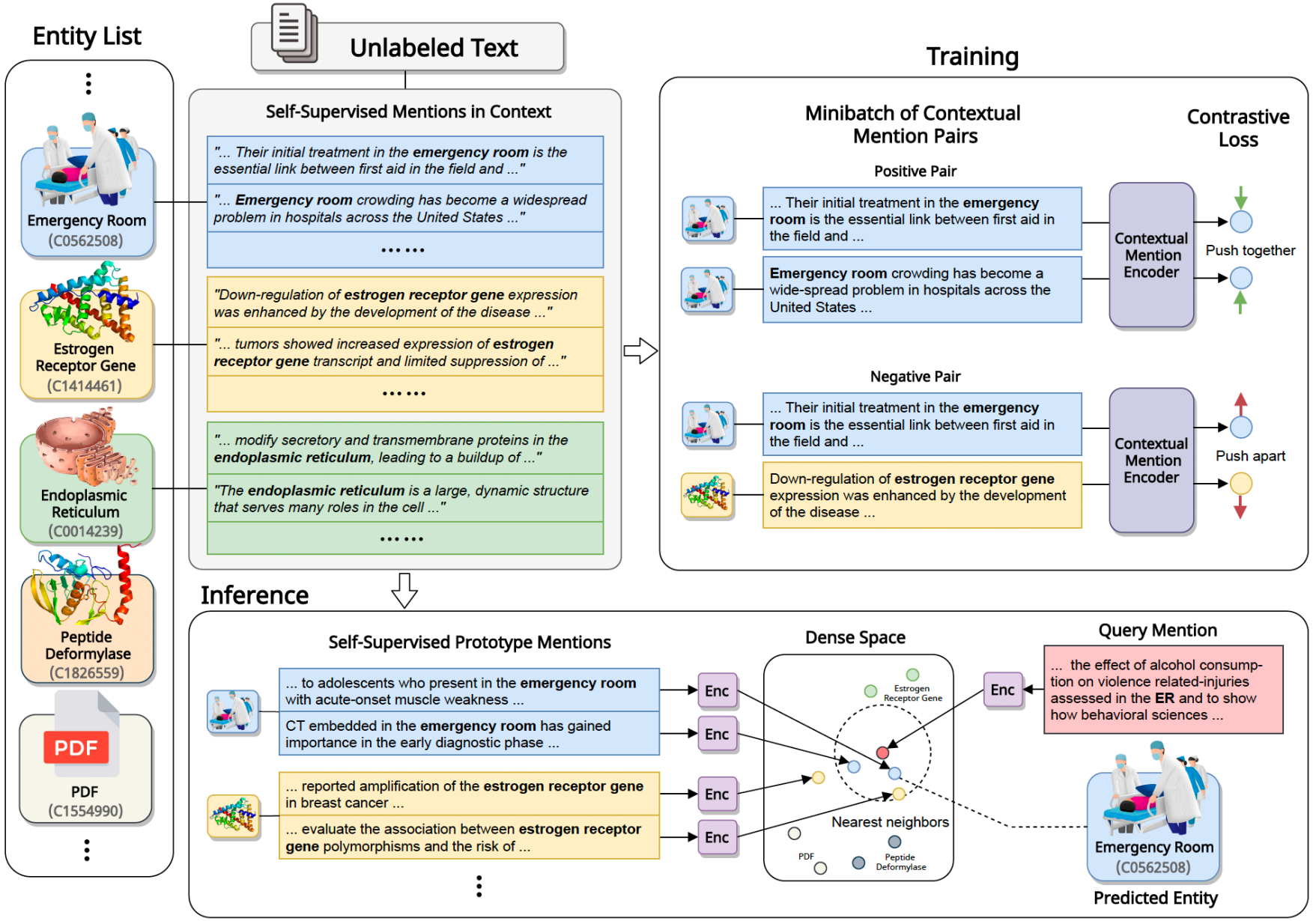


Figure 1: Illustration of knowledge-rich self-supervised entity linking.

- KrissBERT
- Claims SotA performance in 2022, still top or close to top system today
- LLMs not used
- Self-Supervised mentions are identified via exact matching

Hypothesis: Synthetic text mentions generated from large language models can assist with Disease Entity Recognition (DER) and Disease Entity Normalization (DEN)

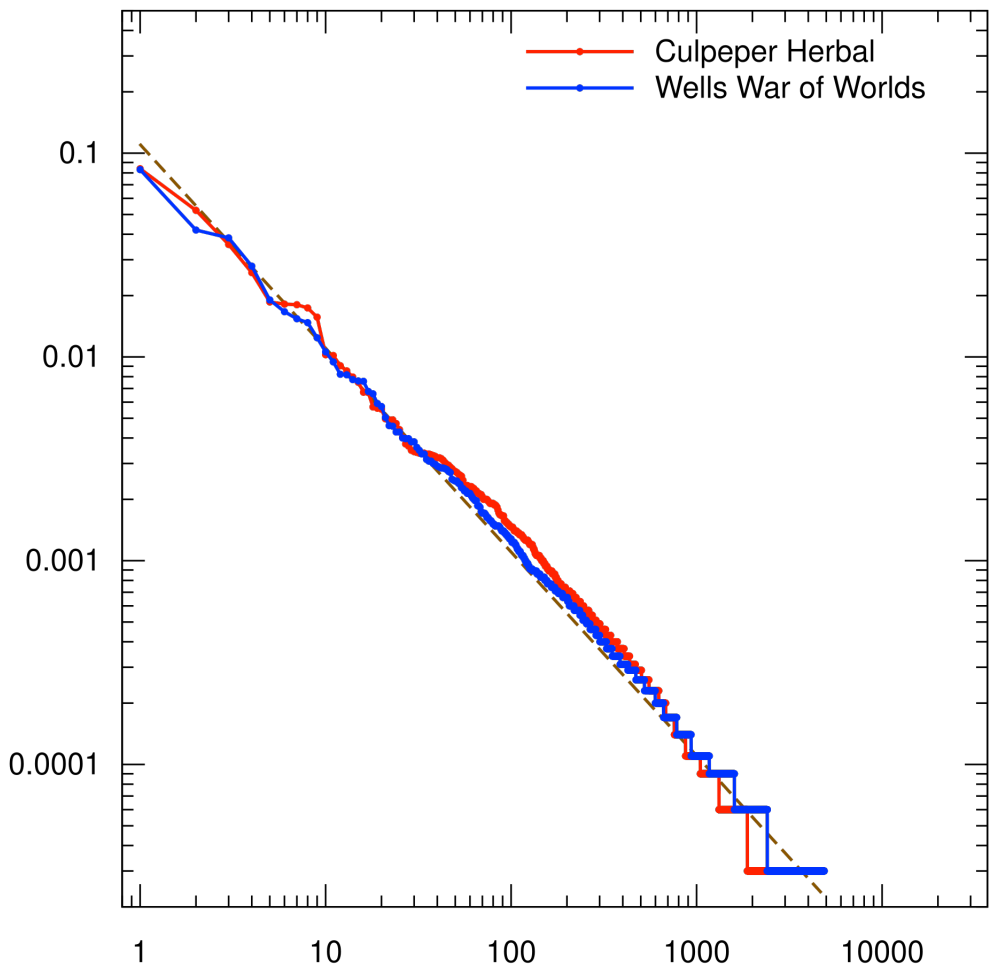
Why disease?

	Disease CUIs			
	With	**Without**	**CUI Average**	**Total**
Synonyms	217252	102129	2.84916	909967
Definitions	53432	265949	0.21725	69417

Table A1 Distribution of Synonyms and Definitions for CUIs in the UMLS 2019AB Disease Semantic Group. There are 319381 Disease Group CUIs.

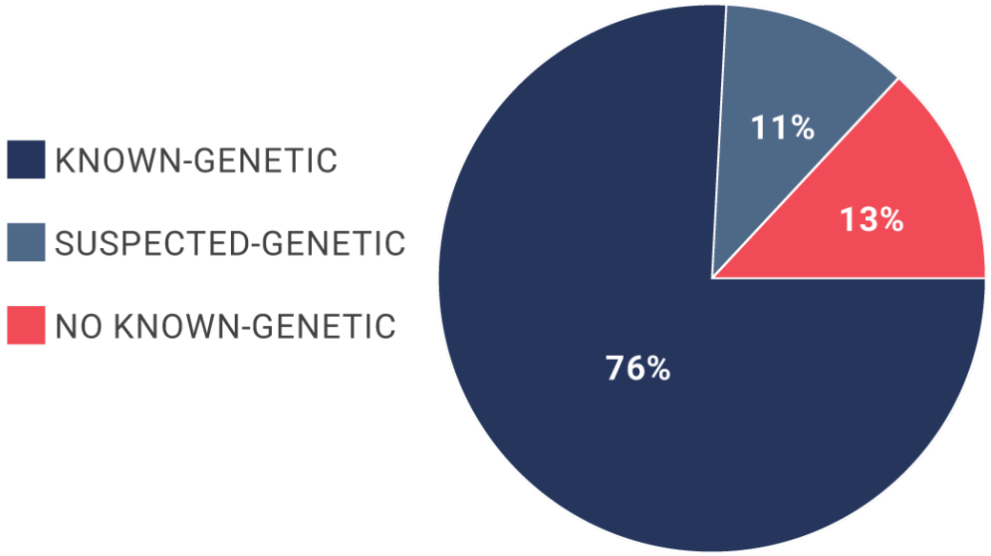
Zipf's Law & Rare Disease

$$\text{word frequency} \propto \frac{1}{\text{word rank}}$$



Rare Disease Occurrence and Relevance

- 7K rare disease affecting 25-40 million Americans

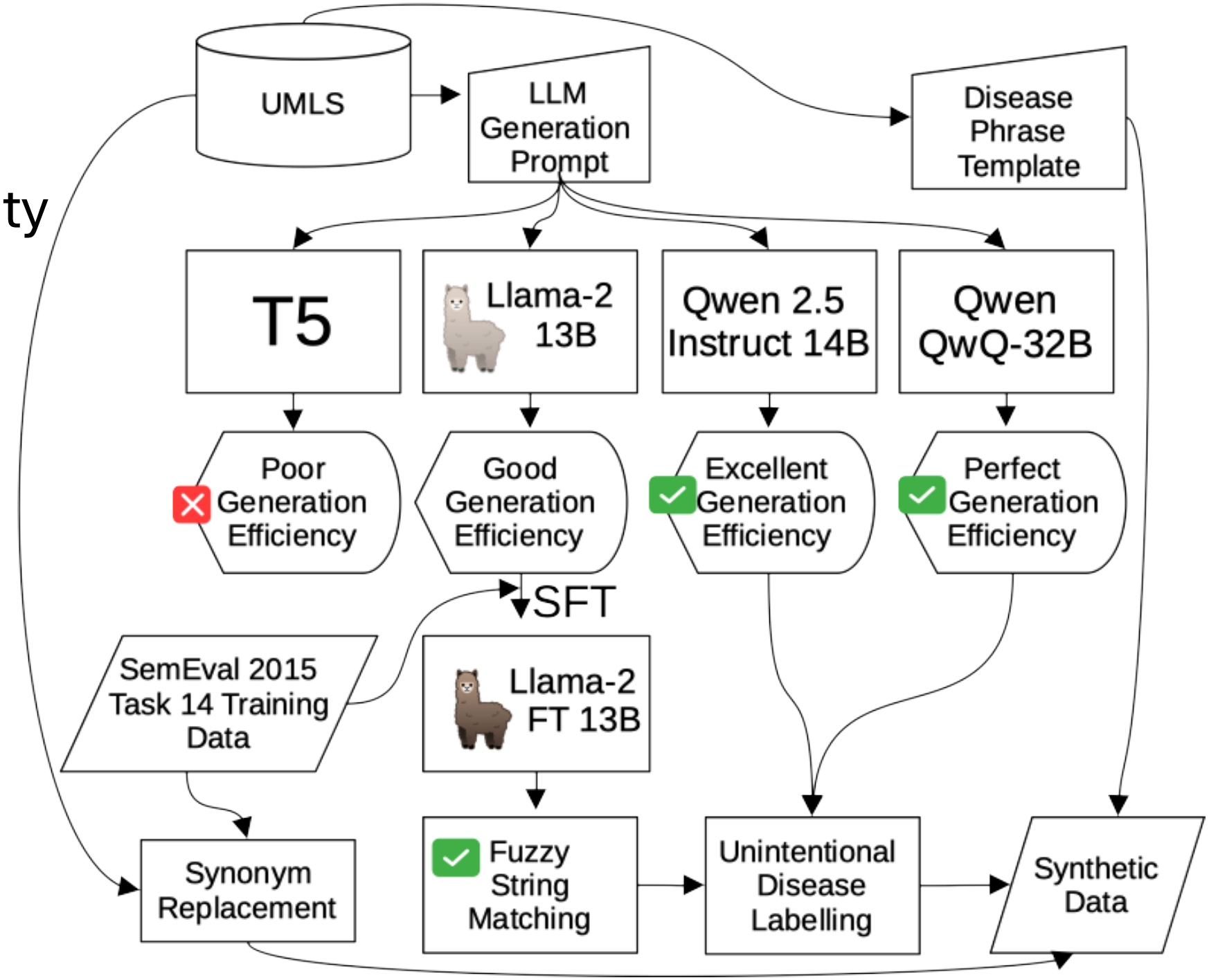


Rare Disease Graphic: https://rare-x.org/case-studies/the-power-of-being-counted/

Zipf's Law Figure: By Jorge Stolfi - Own work, CC BY-SA 4.0, https://commons.wikimedia.org/w/index.php?curid=131702395

Progressive Improvement in Generation Capability

- T5 (not feasible)
- Llama-2 (requires fine-tuning, about 65%)
- Qwen 2.5 Instruct (90%?+)
- QwQ-32B
  - No need to fine-tune or prompt engineer
  - Correct 1CUI tag 100%
  - No concept name mention replacements



Llama-2 Input

SFT Label is SemEval 2015 Task 14 Notes

UMLS Definition?	Prompt
Yes	Pretend you are a physician: Write a clinical note for a patient that mentions the condition MENTION either explicitly or as a synonym or abbreviation to this condition. It is also known as LIST OF NAMES OF MENTION IN UMLS. It is defined as UMLS DEFINITION. Place tokens <1CUI>before and after the mention of this condition. For example <1CUI>MENTION <1CUI>.
No	Pretend you are a physician: Write a clinical note for a patient that mentions the condition MENTION either explicitly or as a synonym or abbreviation to this condition. It is also known as LIST OF NAMES OF MENTION IN UMLS. For example <1CUI>MENTION <1CUI>.

Table 1 Prompts Used in Training LLaMa-2 13B Chat Model. The arbitrarily chosen delimiter <1CUI>is used to enclose the mention and identify the text associated with the CUI.

Llama 2 Output

Disease	Output
Beryllium Disease	Discharge date: [**2013-08-24**] date of birth: [**1947-12-28**] sex: m service: cardiothoracic allergies: patient recorded as having no known allergies to drugs attending:[**attending info 212**] chief complaint: chest pain major surgical or invasive procedure: <1CUI> beryllium dis </1CUI>history of present illness: mr. [**lastname 707**] is a 65-year-old male who presented to the [**hospital1 199**] [**doctor first name 3**] with chest pain he is a former worker at a beryllium plant and was exposed to beryllium.
Metastatic Pancreatic Neuroendocrine Tumor	Discharge date: [**2018-11-03**] date of birth: [**1934-12-02**] sex: m service: medicine allergies: patient recorded as having no known allergies to drugs attending:[**attending info 1699**] chief complaint: h/o <1CUI>metastatic pancreatic neuroendocrine tumor </1CUI>. major surgical or invasive procedure: 1. resection of metastatic pancreatic neuroendocrine tumor.

Table 2 Examples of Generated Disease Mentions (GDMs)

5 examples per CUI

QwQ-32B Generation: C0000921 “Accidental Fall”

Doctor's Note:
Patient Name: Jane Doe
Age/Sex: 68/Female
Chief Complaint: Lateral knee pain after a recent <1CUI> **stumble while ascending stairs** <1CUI>.

Assessment:
Mechanism of injury consistent with a <1CUI> **slip-and-fall episode** <1CUI>, likely leading to contusion and soft tissue injury. Risk factors include prior osteoarthritis and poor lighting in home environment.

Plan:
1. Apply ice and elevate limb.
2. Advise nonsteroidal anti-inflammatory drugs (NSAIDs) for pain.
3. Refer to occupational therapy for home safety evaluation to mitigate future <1CUI> **unsteady gait-related incidents** <1CUI>.
4. Follow-up in 1 week.

How do we use the Synthetic Data?

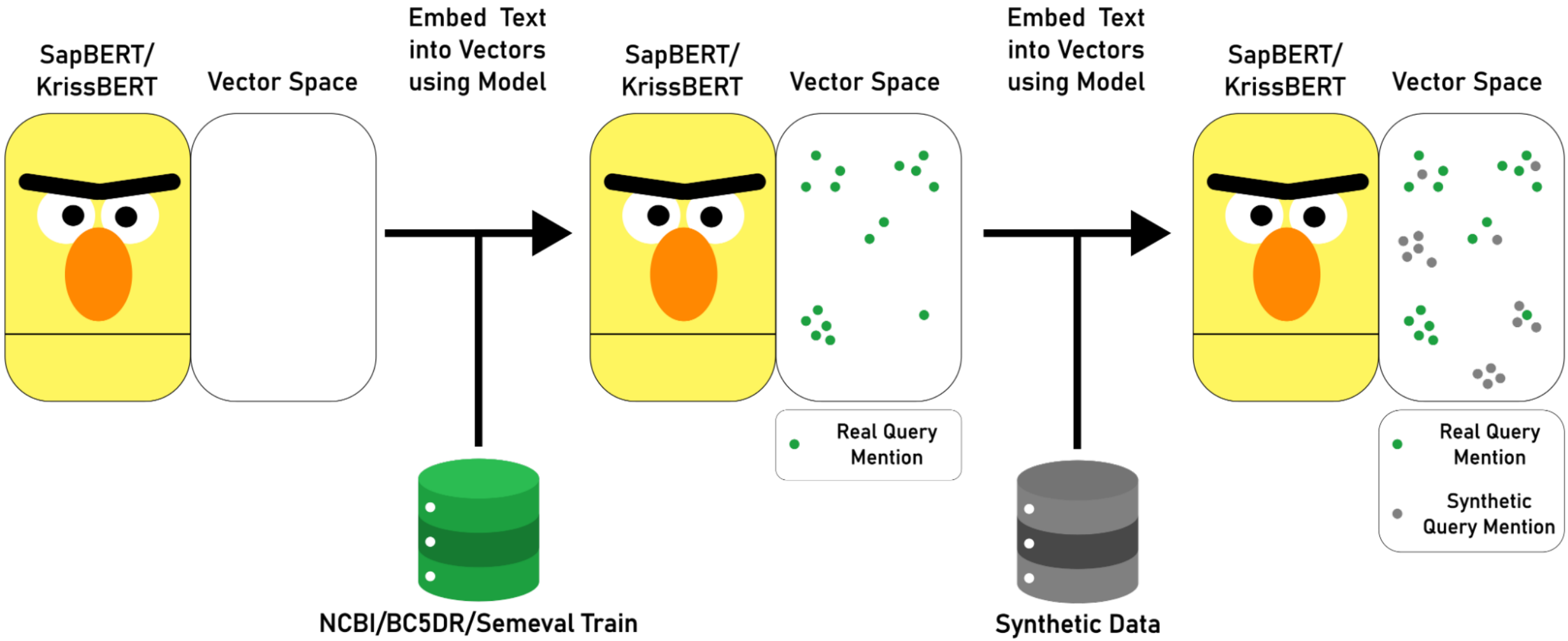
- Naïve
  - Use **ALL** synthetic data
- Ideal
  - Only use synthetic data **IN the TEST split**
- Ablation
  - Only use synthetic data **NOT IN the TEST split**
  - Assess how much performance is coming from generating data in the Test split
- Supplemental
  - Only use synthetic data **NOT in the TRAIN split**
  - Assumes human labels are better

Llama 2 Synthetic Data Augmentation Statistics

		Original		Naive		Ideal		Supp.		Ablation	
Dataset	**Split**	**CUI**	**Mnt.**	**CUI**	**Mnt.**	**CUI**	**Mnt.**	**CUI**	**Mnt.**	**CUI**	**Mnt.**
SemEval	Train	1689	16220	920	128914	212	749	✗	126243	708	128165
	Test	383	1523	250	1523	250	1523	38	1523	✗	1523
BC5DR	Train	634	4318	398	128914	94	1255	✗	127628	304	127658
	Test	196	4135	133	4135	133	4135	39	4135	✗	4135
NCBI	Train	655	5091	435	128914	256	384	✗	127721	179	128528
	Test	639	952	428	952	428	952	172	952	✗	952

Table 3 The total number of disease concepts (CUIs) and mentions for the original dataset are shown under the Original heading. For evaluation of each augmentation strategy, the original held-out test split is used. For training, the total number of mentions (Mnt.) used is shown. The number of concepts (CUIs) from generated mentions that overlap concepts in the the original train and test for that dataset are shown in the CUI columns for each strategy.

Disease Entity Normalization Methodology



Model	Augmentation	Accuracy			OOD Accuracy		
		Top 1	**Top 5**	**Top 50**	**Top 1**	**Top 5**	**Top 50**
SciSpacy	N/A	0.3986	N/A	N/A	N/A	N/A	N/A
QuickUMLS	N/A	0.2703	N/A	N/A	N/A	N/A	N/A
KrissBERT	None (Baseline)	0.7677	0.8263	0.8491	0.0000	0.0000	0.0000
	Naive	**0.8314**	**0.9242**	**0.9531**	**0.4240**	**0.6581**	**0.7128**
	Ideal	0.7203	**0.8659**	**0.9262**	**0.2295**	**0.4514**	**0.6459**
	Supplemental	0.7112	**0.8614**	**0.9249**	**0.2340**	**0.4544**	**0.6474**
	Ablation	0.6765	0.7945	0.8294	0.0000	0.0000	0.0000
SapBERT	None (Baseline)	0.7713	0.7927	0.8297	0.0000	0.0000	0.0000
	Naive	**0.8568**	**0.8866**	**0.9323**	**0.5500**	**0.6112**	**0.6860**
	Ideal	0.7033	**0.8230**	**0.9006**	**0.3667**	**0.4736**	**0.6173**
	Supplemental	0.6959	**0.8154**	**0.8943**	**0.3659**	**0.4736**	**0.6173**
	Ablation	0.6595	0.7260	0.7746	0.0000	0.0000	0.0000

Table 6 BC5DR Synthetic Normalization Results. **Bold** is a model that beats the baseline, *italics* means the model beat the baseline at all thresholds.

Qwen 2.5 Instruct Generation Negative Controls (LLM-Free)

Training Data Synonyms

- C0600228,Lidocaine-induced <1CUI> **arrest cardio respiratory** </1CUI> . Intravenous administration of a single 50-mg bolus of lidocaine in a 67-year-old man resulted in profound depression of the activity of the sinoatrial and atrioventricular nodal pacemakers.
- C0600228,Lidocaine-induced <1CUI> **Cardiopulmonary arrest** </1CUI> . Intravenous administration of a single 50-mg bolus of lidocaine in a 67-year-old man resulted in profound depression of the activity of the sinoatrial and atrioventricular nodal pacemakers.

Templated Generation

Patient Name: [Patient Name 2301]

Chief Complaint: Symptoms related to <1CUI>**{disease}</**1CUI>

History of Present Illness: Patient presents with signs and symptoms consistent with <1CUI>**{disease}</**1CUI>. Further evaluation and diagnostic workup are in progress to confirm severity and appropriate management.

- Plan: Proceed with necessary assessments and initiate appropriate care as indicated.

Model	Augmentation	Accuracy			OOD Accuracy		
		Top 1	Top 5	Top 50	Top 1	Top 5	Top 50
SciSpacy	N/A	0.3986	N/A	N/A	N/A	N/A	N/A
QuickUMLS	N/A	0.2703	N/A	N/A	N/A	N/A	N/A
KrissBERT	None (Baseline)	0.7604	0.8502	0.8705	0.0000	0.0000	0.0000
	Ablation (LLM)	0.7354	0.7906	0.8121	0.0000	0.0000	0.0000
	Supplemental (LLM)	0.8182	0.9132	0.9663	**1.0000**	**1.0000**	**1.0000**
	Naive (Synonym)	0.7171	0.7835	0.8100	0.0000	0.0000	0.0000
	Naive (Template)	0.7477	0.8182	0.8856	0.6667	0.6667	**1.0000**
	Naive (LLM)	0.8212	**0.9173**	**0.9673**	**1.0000**	**1.0000**	**1.0000**
	Ideal (Synonym)	0.7457	0.7937	0.8131	0.0000	0.0000	0.0000
	Ideal (Template)	0.7967	0.8897	0.9520	0.6667	**1.0000**	**1.0000**
	Ideal (LLM)	**0.8264**	**0.9173**	**0.9673**	**1.0000**	**1.0000**	**1.0000**
SapBERT	None (Baseline)	0.7276	0.7582	0.7862	0.0000	0.0000	0.0000
	Ablation (LLM)	0.7270	0.7518	0.7809	0.0000	0.0000	0.0000
	Supplemental (LLM)	0.8215	**0.8633**	**0.9029**	**0.5000**	**0.5760**	**0.6462**
	Naive (Synonym)	0.7181	0.7455	0.7693	0.0000	0.0000	0.0000
	Naive (Template)	0.7202	0.7899	0.8569	0.2456	0.3275	0.4503
	Naive (LLM)	**0.8268**	**0.8601**	**0.9060**	**0.5000**	0.5256	**0.6462**
	Ideal (Synonym)	0.7365	0.7582	0.7746	0.0000	0.0000	0.0000
	Ideal (Template)	0.8094	0.8574	0.9065	0.4152	0.5088	0.5614
	Ideal (LLM)	0.8194	**0.8675**	**0.9071**	**0.5000**	0.5702	**0.6462**

Table 7: NCBI-Disease Qwen 2.5-Instruct Synthetic Normalization Results. **Bold** is a model that beats the baseline, *italics* means the model beat the baseline at all thresholds.

Why Do Mini-Documents with Mentions Help?

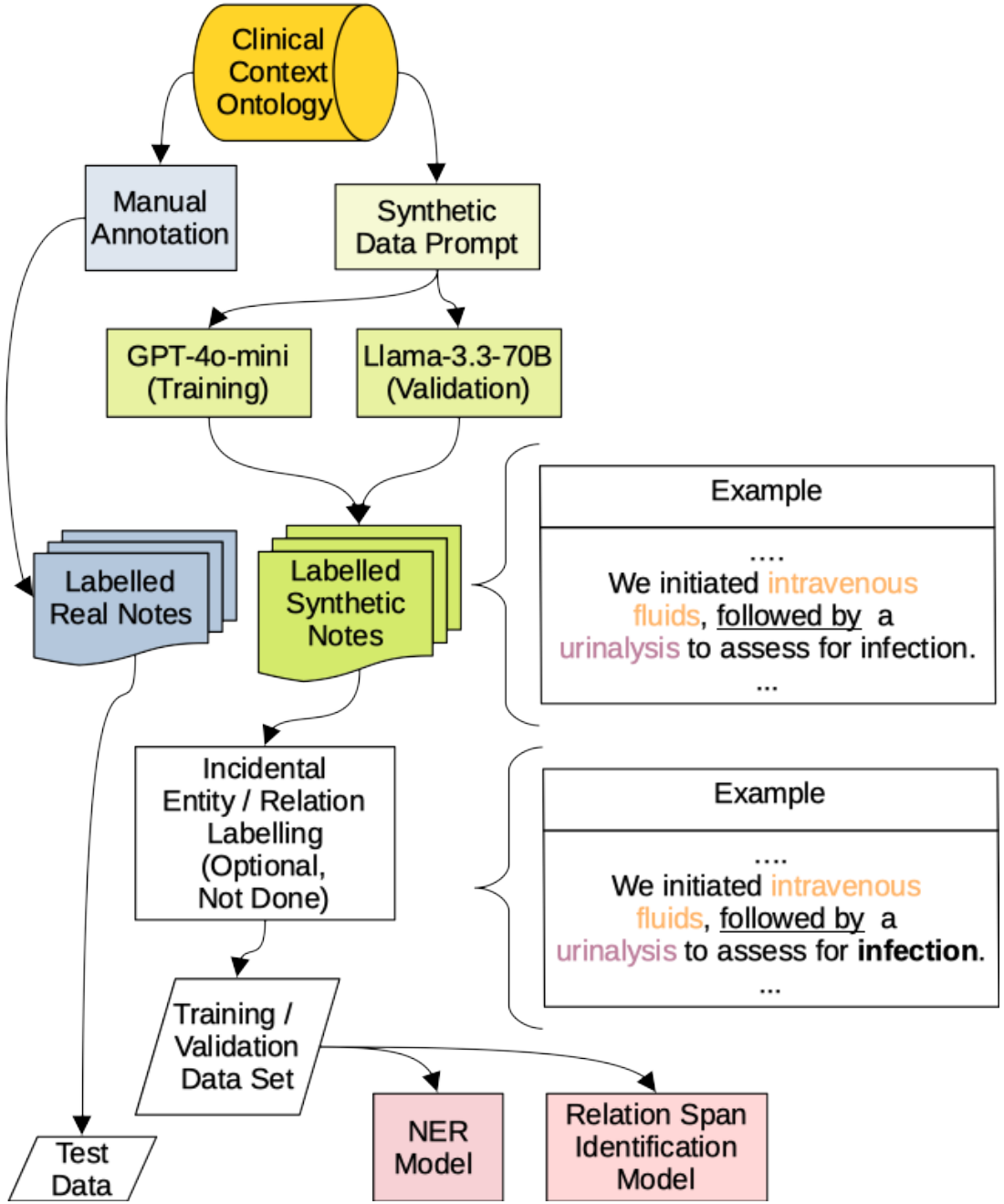
	Disease CUIs			
	With	**Without**	**CUI Average**	**Total**
Synonyms	217252	102129	2.84916	909967
Definitions	53432	265949	0.21725	69417

Table A1 Distribution of Synonyms and Definitions for CUIs in the UMLS 2019AB Disease Semantic Group. There are 319381 Disease Group CUIs.

- Knowledge Graph / UMLS is not always that knowledgeable!
  - Supplementation with Definitions / Synonyms is not enough
  - LLM pre-training sources are vast but needs to be effectively utilized

Future Work

- Relation Extraction
- Human Phenotype Ontology (HPO) Evaluation on multiple data sets
- Additional Entities
- New LLM Evaluations
  - Novel LLM extraction methods



NVidia GPU Grant Program, ”Improving Information Extraction in Biomedical Text”

Acknowledgements

BRATSynthetic

- Tobias O’Leary, Chris Coffee, Andrew Trotter
- Abdulateef Almudaifer, Akhil Nadimpalli
- Luis Mansilla Gonzalez, Salma Aly, Richard E. Kennedy

LLM Generated Synthetic Data

- Kuleen Sasse, Shinjitha Valkakonda
- Richard Kennedy
- Shan Chen, Maio Danila
- Vijay Jain (Qwen 32B HPO Terms)
- Kaiwen He (GPT-4o-mini generated relations)

NIH NIAMS grant P30AR072583,”Building and InnovatinG: Digital heAlth Technology and Analytics”.

NIH NIA grant R01AG057684, ”In Silico Screening of Medications for Slowing Alzheimer’s Disease Progression

Questions?