Stockholm University

Leveraging LLM-created synthetic data for NLP

A new approach to privacy preservation

Department of Computer and Systems Sciences, Stockholm University

Thomas Vakili (thomas.vakili@dsv.su.se)

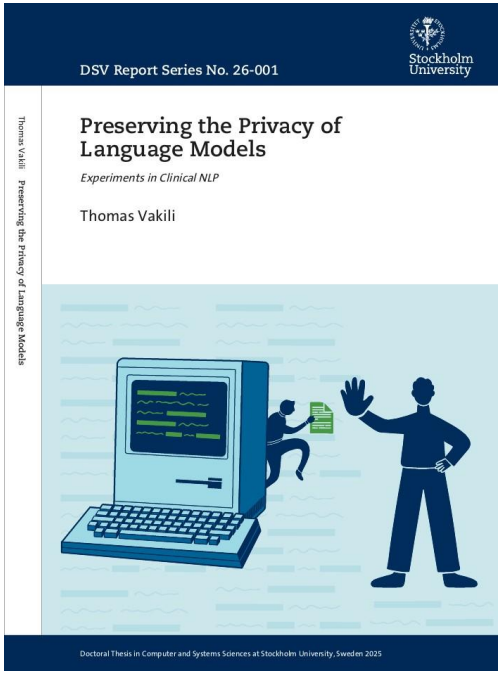
Who am I?

- PhD student since 2021 @Stockholm University
- “Privacy-preserving NLP”
- Previously
  - Engineering @KTH
  - Developer in industry

Thomas Vakili - thomas.vakili@dsv.su.se

Who am I?

- PhD student since 2021 @Stockholm University
- Thesis defence Jan 13!
  - 13:30 in Kista



https://vakili.science/documents/thesis-draft.pdf

The Health Bank

- Data from the Karolinska University Hospital
- Enormous data source
  - More than 18 GB of raw text
  - Millions of electronic health records
- Many different kinds of data, such as:
  - Diagnosis codes
  - Lab measurements
  - ***Text from electronic health records***

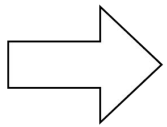
Thomas Vakili - thomas.vakili@dsv.su.se

A decade of building resources

Many task-specific datasets have been built over the years:

- Stockholm EPR **Gastro ICD-10 Corpus**
- Stockholm EPR **PHI Corpus**
- Stockholm EPR **Clinical Entity Corpus**
- Stockholm EPR **Diagnosis Factuality Corpus**
- Stockholm EPR **Adverse Drug Event ICD-10 Corpus**

Pat arrives to hospital with broken tibia.
Anaesthetic given by nurse Andersson.
Paracetamol prescribed by dr Modig.



S82.1

Fracture of upper end of tibia

Stockholm University

NLP & Privacy

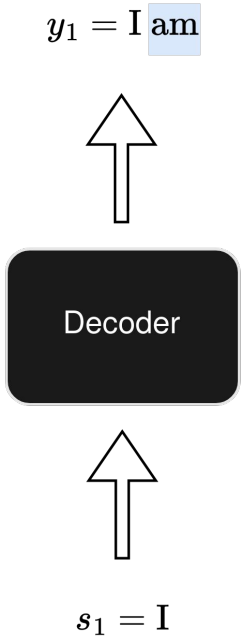
Language Models

- Language modelling is traditionally done "left-to-right"
- GPT-{1, 2, 3, 4, 5?} are generative language models

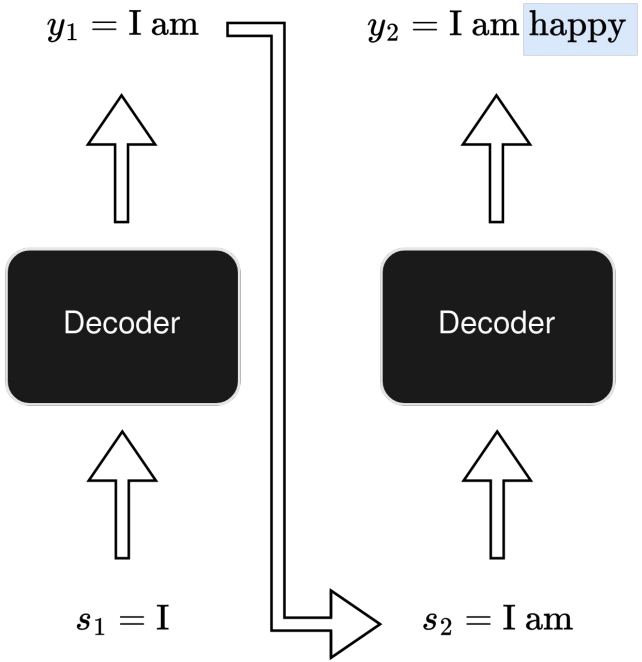
Thomas Vakili is a PhD student at **DSV**

$$x_{i+1} = \underset{w \in V}{\operatorname{argmax}} P(w|x_1, x_2, ..., x_i)$$

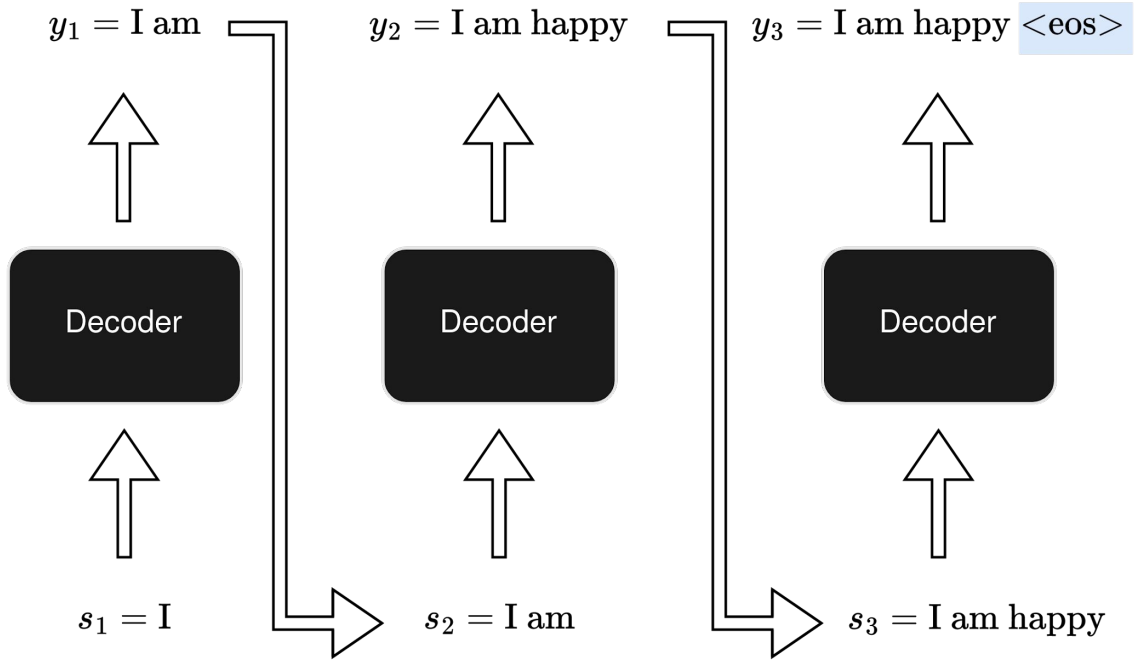
Generative LLMs



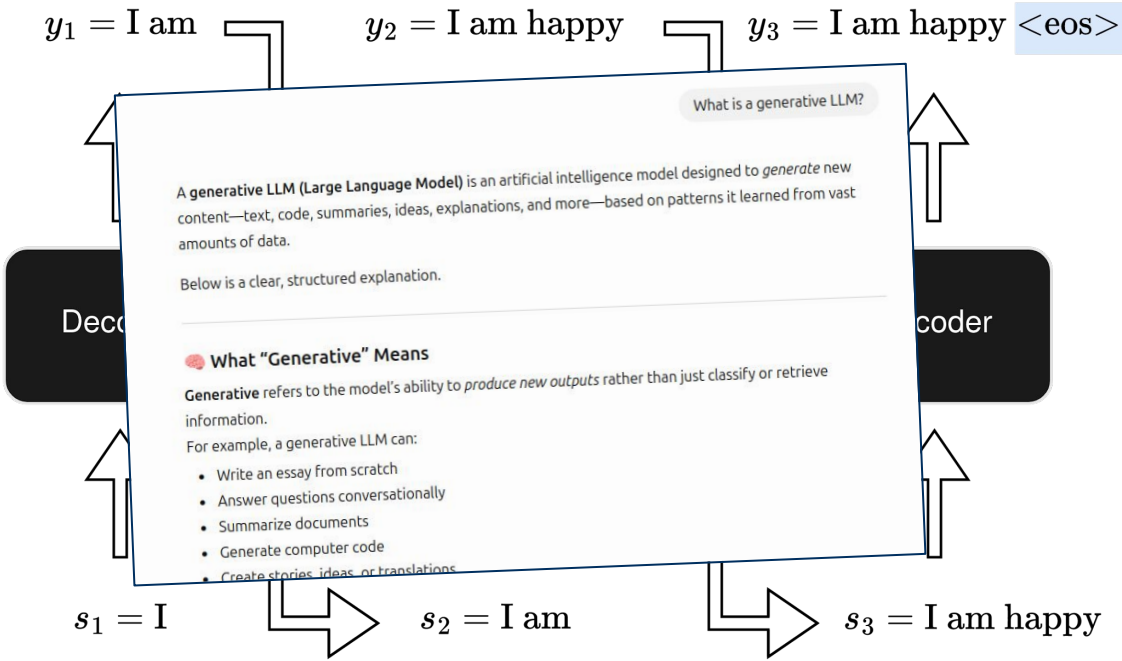
Generative LLMs



Generative LLMs

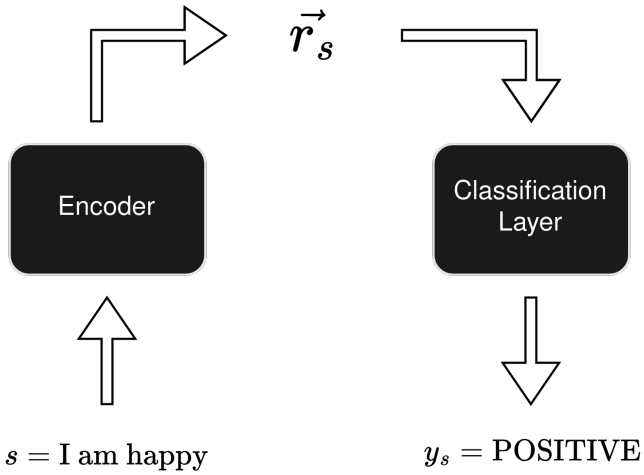


Generative LLMs



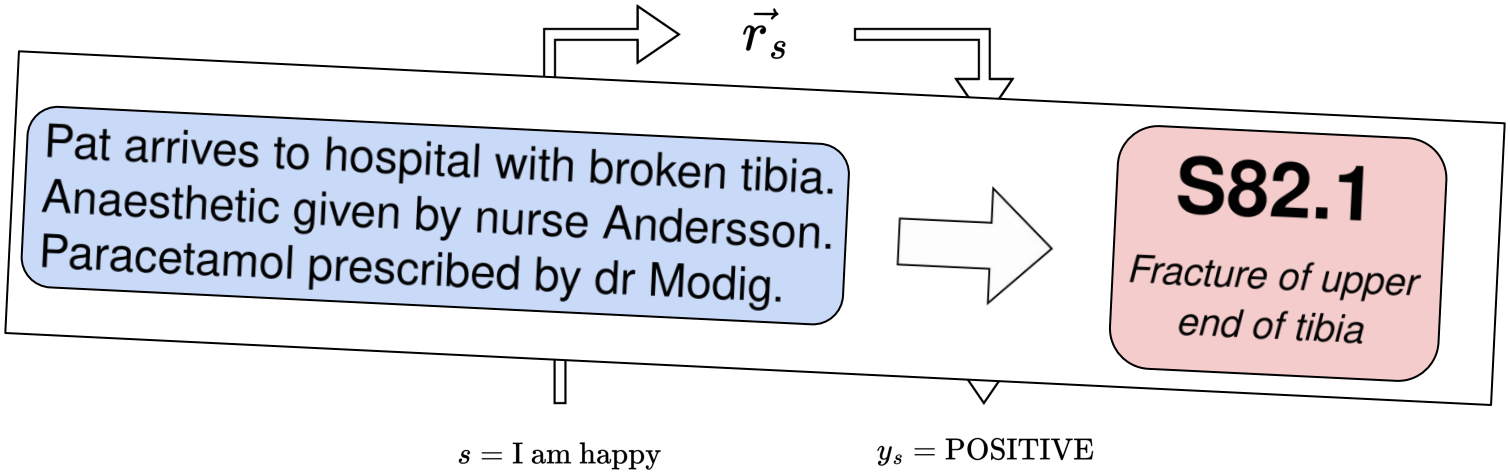
Encoder models

Encode input into vector representations, then classify



Encoder models

Encode input into vector representations, then classify



Thomas Vakili - thomas.vakili@dsv.su.se

Memorization

- LLMs are trained to reproduce their training data
- This sometimes leads to *memorization*

The patient fractured his **[MASK]**

↓

The patient fractured his **femur**

Thomas Vakili - thomas.vakili@dsv.su.se

Memorization

- LLMs are trained to reproduce their training data
- This sometimes leads to *memorization*

The patient's personal number is **[MASK]**

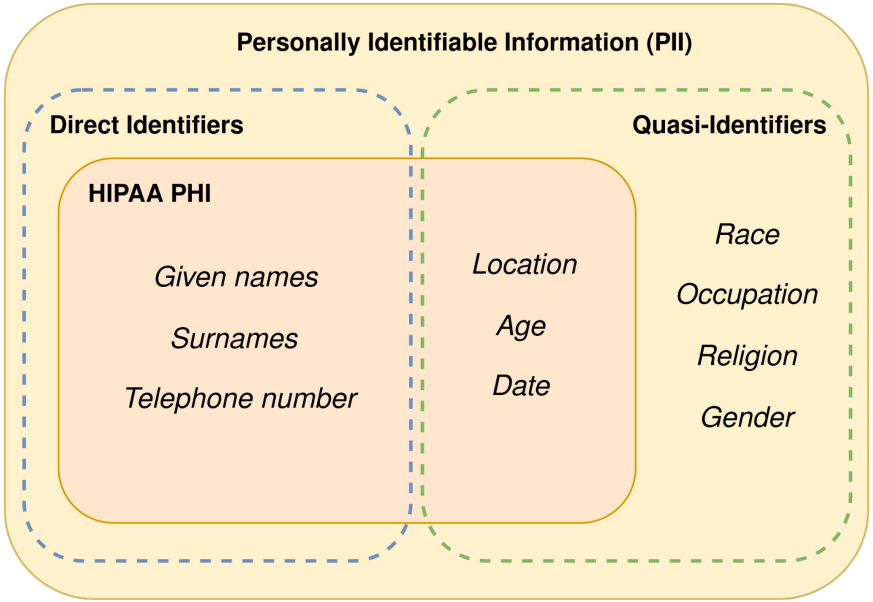
↓

The patient's personal number is **950208-1234**

Thomas Vakili - thomas.vakili@dsv.su.se

PII – Personally Identifiable Information

- PII are everywhere
- Public data ≠ Free to use
- Why was it created?
- Probably not to train LLMs
- Definitely not EHRs



Privacy Attacks

- Training data extraction
  - “What is the most likely output?”
  - Requires finding an effective way to sample
  - How do we know if sample is *real?* Typically using …

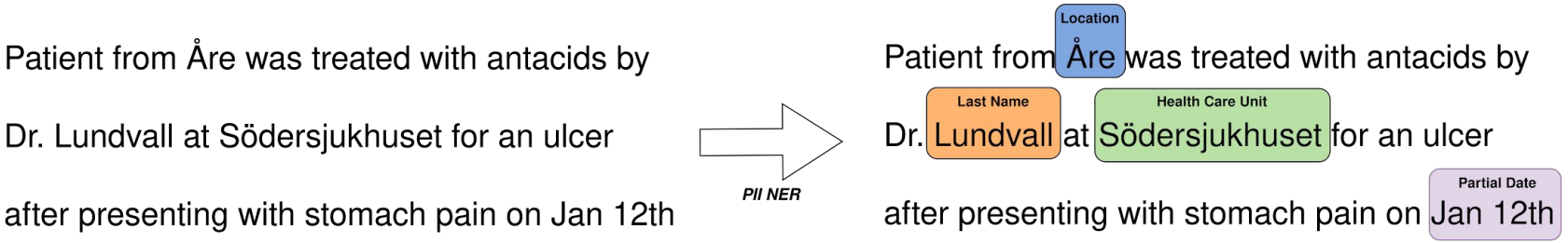
- Membership inference
  - “Has the model seen this datapoint?”
  - Look at how a model *reacts* to a datapoint
  - Proposed for quantifying privacy risks

Privacy-preserving techniques

- Model-oriented
  - Differentially private learning
- Data-oriented
  - Automatic de-identification
  - Synthetic data

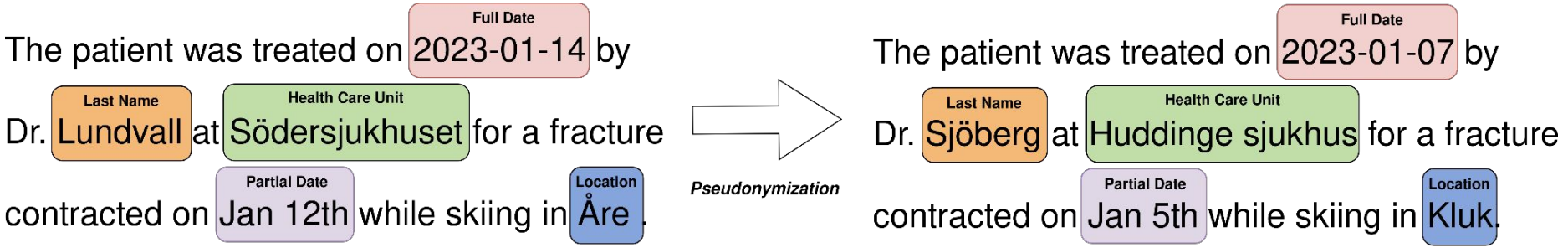
PII identification (NER)

Find tokens representing **personally identifiable information**. The task covers **nine classes** such as **first names, dates, health care units, and locations.**



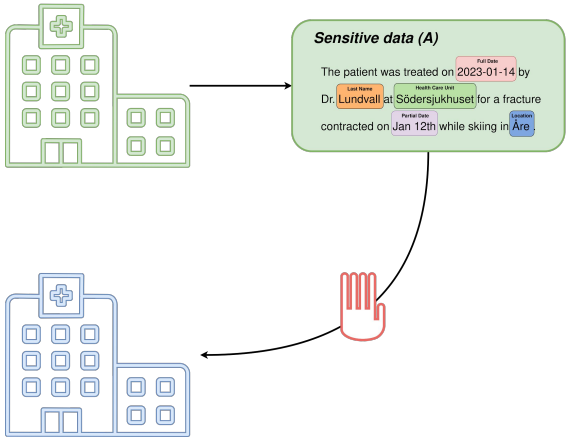
Dalianis, H., & Velupillai, S. (2010). De-identifying Swedish clinical text—Refinement of a gold standard and experiments with Conditional random fields. *Journal of Biomedical Semantics*, *1*(1), 6.

Pseudonymization



The problem

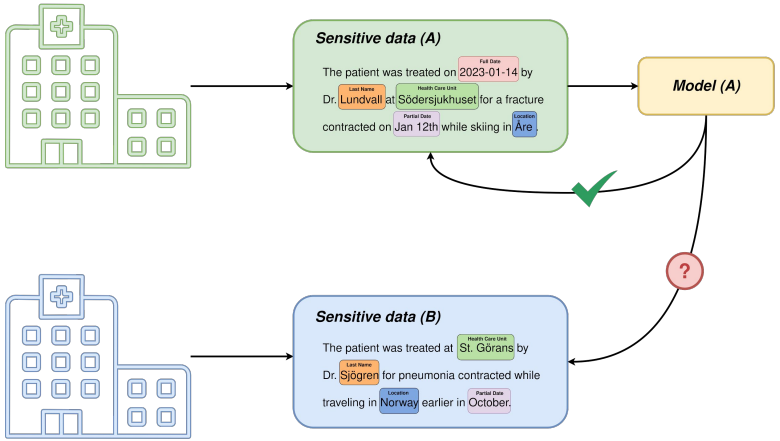
- NLP is increasingly being used in sensitive domains
- Domains like the clinical domain suffer from
  - Data scarcity
  - Difficulties sharing data



21

The problem

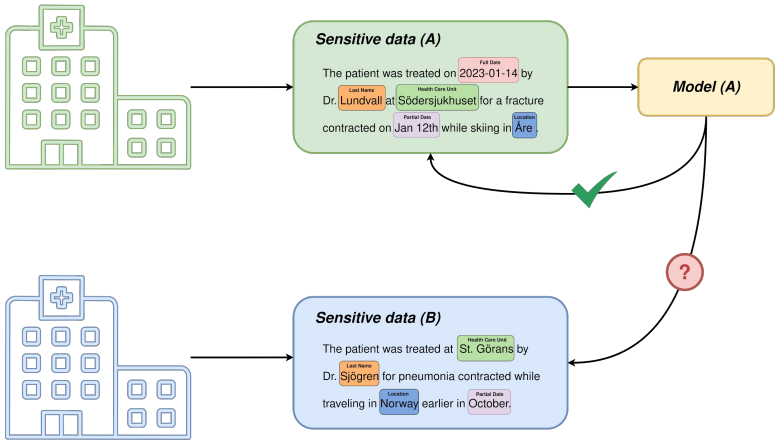
- NLP is increasingly being used in sensitive domains
- Domains like the clinical domain suffer from
  - Data scarcity
  - Difficulties sharing data



22
Thomas Vakili - thomas.vakili@dsv.su.se

The problem

- NLP is increasingly being used in sensitive domains
- Domains like the clinical domain suffer from
  - Data scarcity
  - Difficulties sharing data



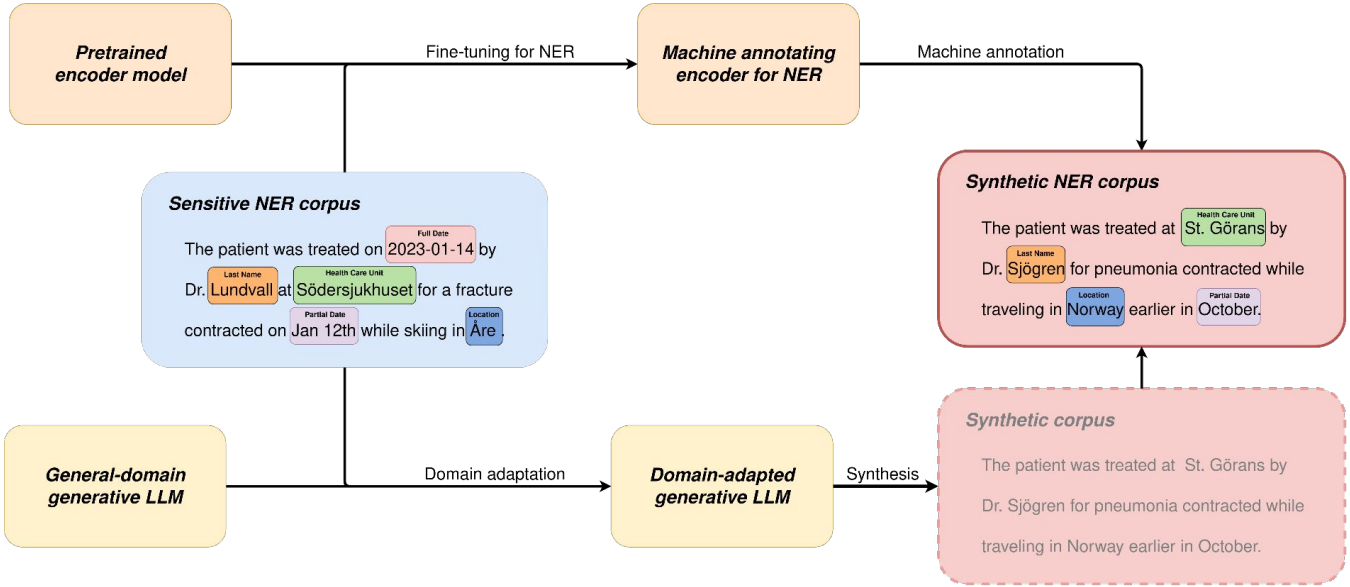
Can we use synthetic data instead?

23

Stockholm University

Experiments

Two approaches



Thomas Vakili, Aron Henriksson, and Hercules Dalianis. 2025. Data-Constrained Synthesis of Training Data for De-Identification. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*.

Two approaches

Below is an instruction that describes a task, paired with an input that provides further context. Write a response that appropriately completes the request.

Instruction:

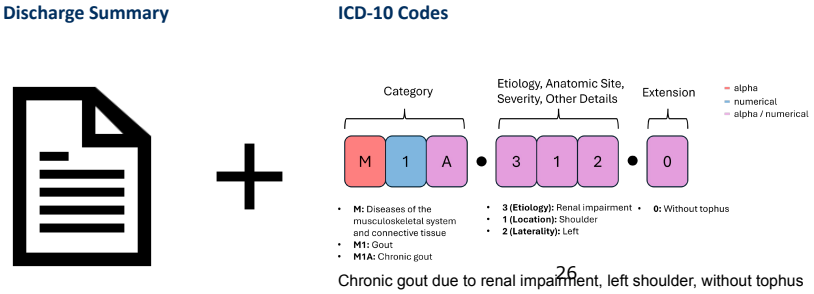
Given a list of textual descriptions of procedure and diagnosis codes, generate a corresponding clinical discharge summary that provides comprehensive and relevant details about the patient's medical history, current condition and treatment received at the hospital.

Input:

["STEMI, unspecified site", "Primary hypertension", "Unspecified hyperlipidemia", "Coronary artery stent placement, drug-eluting", "Cardiac output monitoring", "IV thrombolytic administration"]

Response:

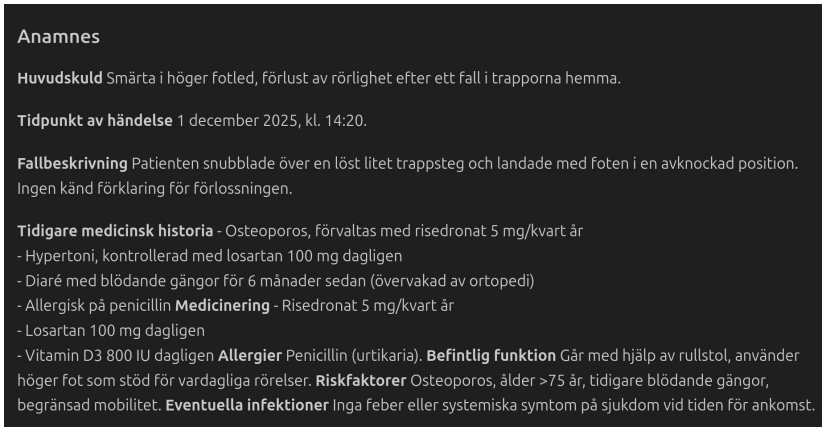
Discharge Summary



Lotta Kiefer, Jesujoba O Alabi, Thomas Vakili, Hercules Dalianis, Dietrich Klakow. 2025. Data-Constrained Synthesis of Training Data for De-Identification. In *Proceeedings of the 15th International Conference on Recent Advances in Natural Language Processing.*

(Domain adaptation)

LLMs are trained on internet data — general domain data



Anamnes

Huvudskuld Smärta i höger fotled, förlust av rörlighet efter ett fall i trapporna hemma.

Tidpunkt av händelse 1 december 2025, kl. 14:20.

Fallbeskrivning Patienten snubblade över en löst litet trappsteg och landade med foten i en avknockad position. Ingen känd förklaring för förlossningen.

Tidigare medicinsk historia - Osteoporos, förvaltas med risedronat 5 mg/kvart år
- Hypertoni, kontrollerad med losartan 100 mg dagligen
- Diaré med blödande gängor för 6 månader sedan (övervakad av ortopedi)
- Allergisk på penicillin **Medicinering** - Risedronat 5 mg/kvart år
- Losartan 100 mg dagligen
- Vitamin D3 800 IU dagligen **Allergier** Penicillin (urtikaria). **Befintlig funktion** Går med hjälp av rullstol, använder höger fot som stöd för vardagliga rörelser. **Riskfaktorer** Osteoporos, ålder >75 år, tidigare blödande gängor, begränsad mobilitet. **Eventuella infektioner** Inga feber eller systemiska symtom på sjukdom vid tiden för ankomst.

We can tune them with clinical data to become more realistic!

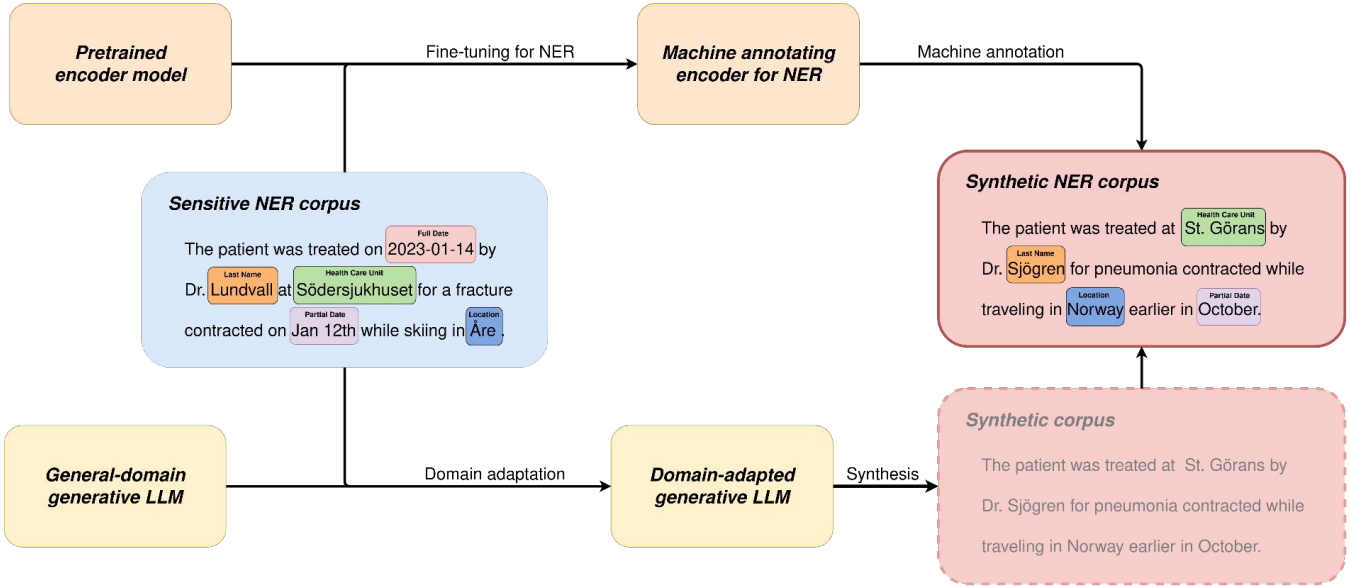
27

Vakili et al. (2025)

- LLMs can generate convincing synthetic text – nowadays
- Previous studies into using synthetic data
  - Assess feasibility of training using synthetic data
  - Study privacy risks
- *We systematically study the conditions for success*
- Data: NER corpora for PII identification

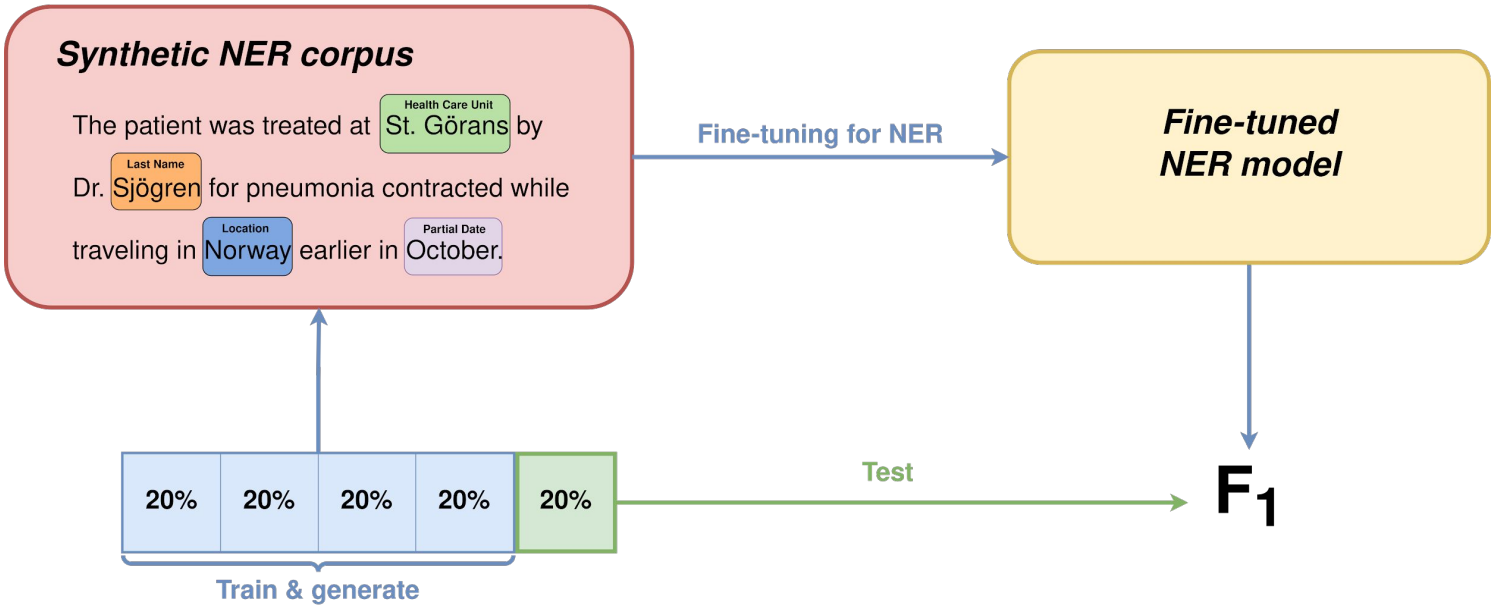
28
Thomas Vakili - thomas.vakili@dsv.su.se

Generate and machine annotate



Thomas Vakili - thomas.vakili@dsv.su.se

Evaluate downstream utility



30

Datasets

- MEDDOCAN [1]
  - 1,000 documents in Spanish
  - 19 different classes of PII
- SEPR PHI: Stockholm EPR PHI Pseudo Corpus [2]
  - 21,553 short documents in Swedish
  - 9 different classes of PII

31

Language Models

- Generative models
  - FLOR (6.3B & 1.3B) for Spanish [3]
  - GPT-SW3 (6.7B & 1.3B) for Swedish [4]
- Encoder models
  - roberta-base-bne for Spanish [5]
  - SweDeClin-BERT for Swedish [6]

32

Measuring data quality

- *Utility:* Downstream token-level F_1
- *Privacy*
  - 5-gram overlap
  - **Sensitive** 5-gram overlap

33

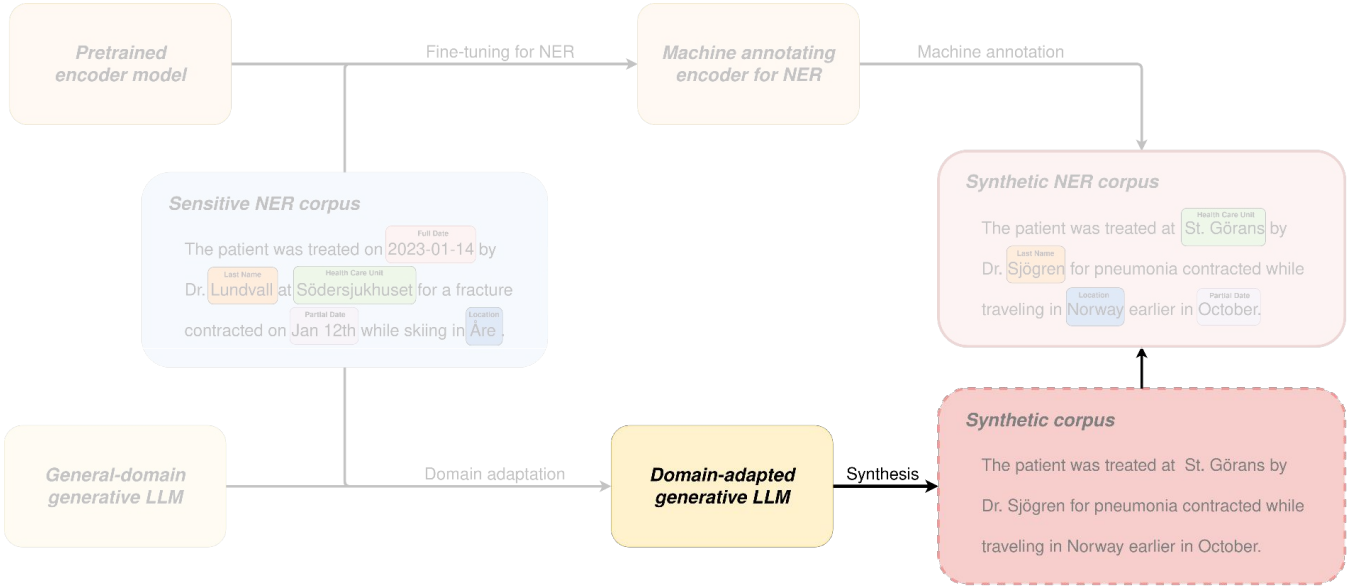
Thomas Vakili - thomas.vakili@dsv.su.se

Measuring data quality

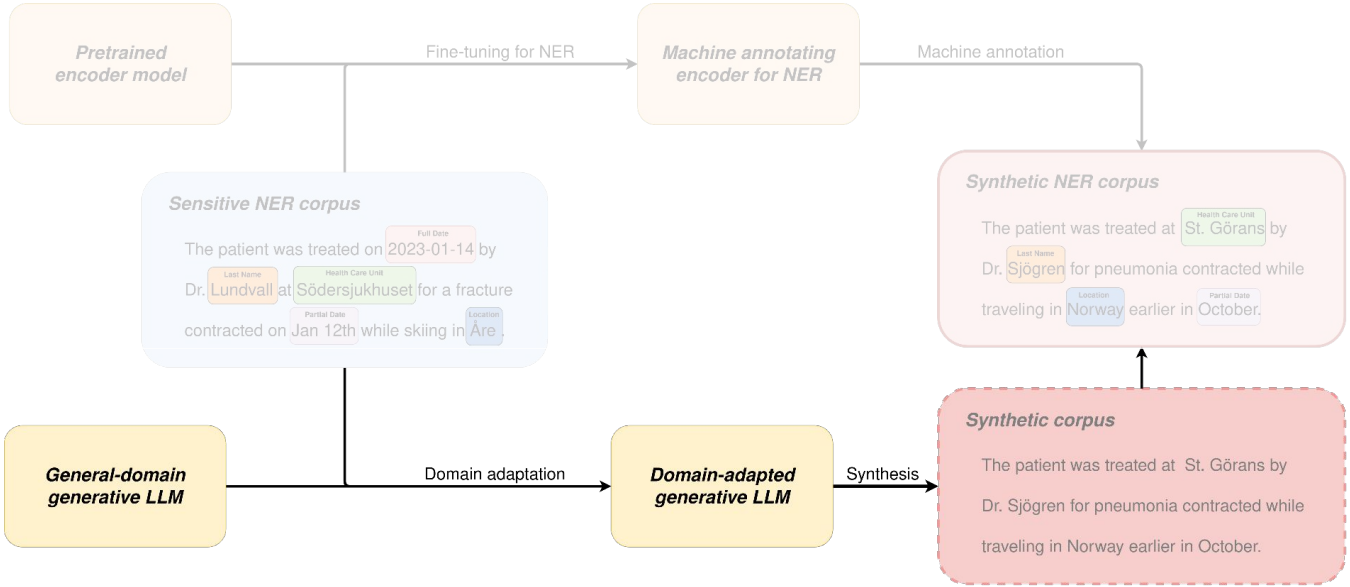
- *Utility:* Downstream token-level F_1
- *Privacy*
  - 5-gram overlap
  - **Sensitive** 5-gram overlap
- *Descriptive metrics*
  - Lexical diversity
  - Document length
  - Number of entities per document

34
Thomas Vakili - thomas.vakili@dsv.su.se

1: Size of generated corpus

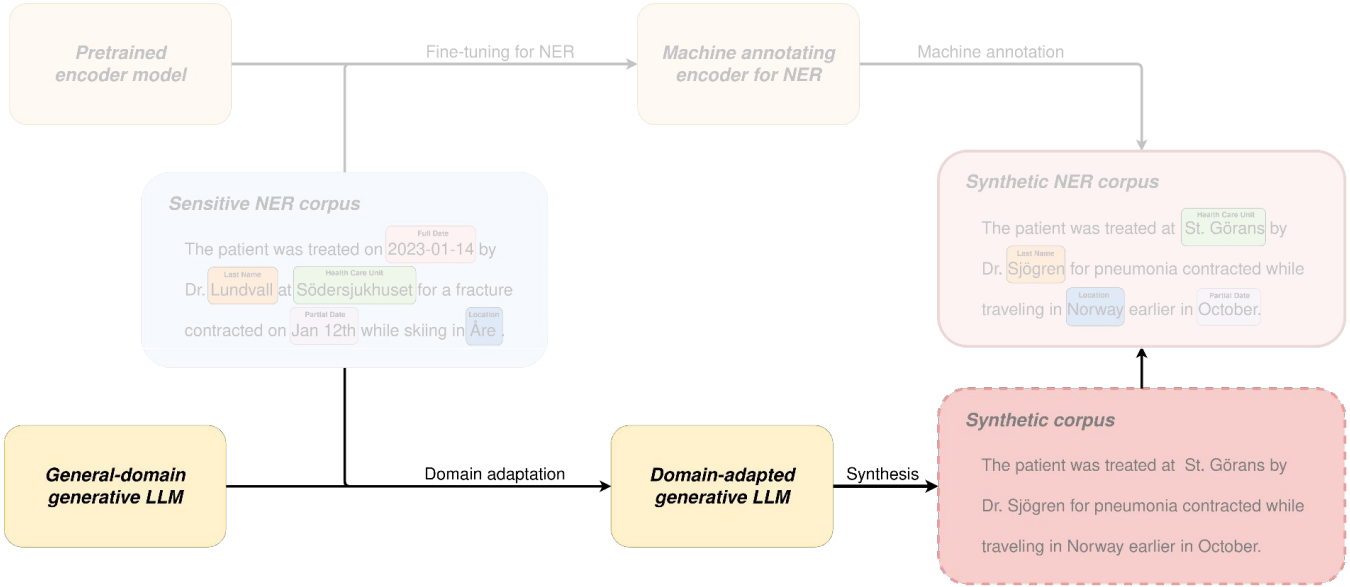


2: Size of generative LLM

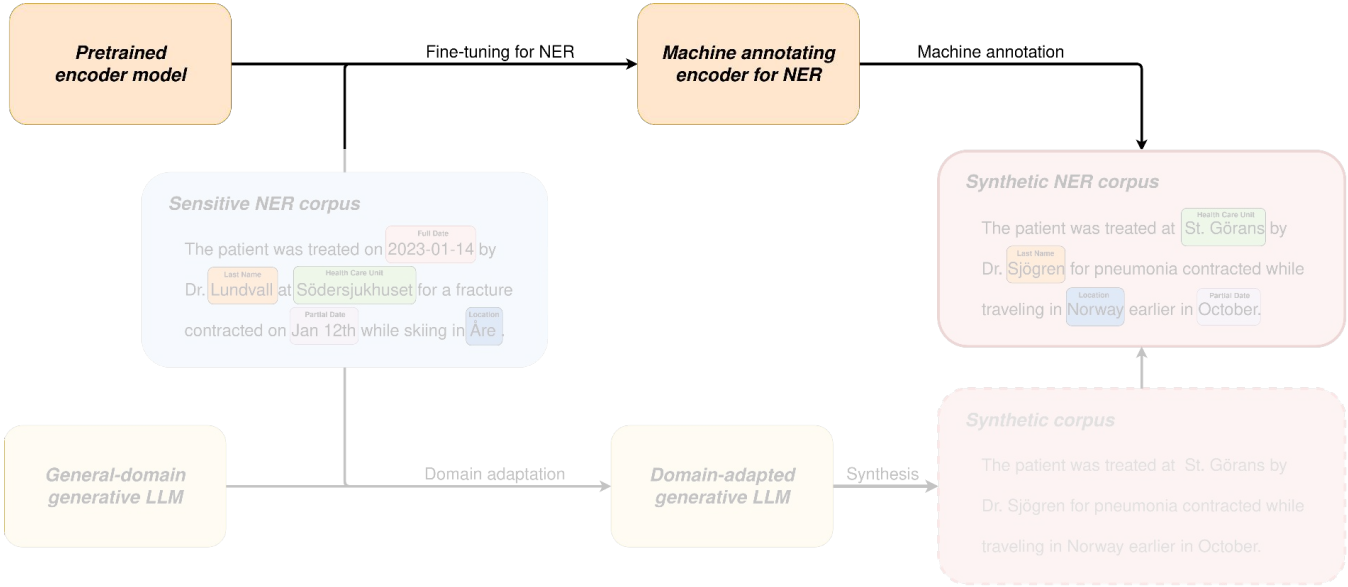


Thomas Vakili - thomas.vakili@dsv.su.se

3: Data for domain adaptation



4: Data for machine annotator



Thomas Vakili - thomas.vakili@dsv.su.se

Does it work?

Not so much impact from LLM size

Model size	SEPR PHI	MEDDOCAN
Small	0.883 ± 0.006	0.973 ± 0.004
Larger	0.896 ± 0.007	0.973 ± 0.003
Gold	0.926 ± 0.005	0.973 ± 0.004

Does it work?

No big difference when 1x or 4x of original size

Synthesized amount	SEPR PHI	MEDDOCAN
5%	0.814 ± 0.008	0.938 ± 0.006
100%	0.889 ± 0.009	0.968 ± 0.005
400%	0.896 ± 0.007	0.973 ± 0.004
Gold	0.920 ± 0.008	0.977 ± 0.005

Does it work?

Diminishing returns from **d**omain **a**daptation (!)

% for d.a.	SEPR PHI	MEDDOCAN
0%	0.547 ± 0.178	0.295 ± 0.011
5%	0.873 ± 0.014	0.313 ± 0.032
25%	0.877 ± 0.010	0.970 ± 0.005
50%	0.896 ± 0.007	0.970 ± 0.005
95%	0.896 ± 0.007	0.973 ± 0.003
Gold	0.926 ± 0.005	0.978 ± 0.005

Does it work?

Diminishing returns from **d**omain **a**daptation (!)

% for d.a.	SEPR PHI	MEDDOCAN
0%	0.547 ± 0.178	0.295 ± 0.011
5%	0.873 ± 0.014	0.313 ± 0.032
25%	0.877 ± 0.010	0.970 ± 0.005
50%	0.896 ± 0.007	0.970 ± 0.005
95%	0.896 ± 0.007	0.973 ± 0.003
Gold	*0.926 ± 0.005*	*0.978 ± 0.005*

Does it work?

Instead, **m**achine **a**nnotation seems to matter more

% for m.a.	SEPR PHI		MEDDOCAN	
	Gold	Synthetic	Gold	Synthetic
5%	0.707 ± 0.037	0.725 ± 0.039	0.931 ± 0.012	0.942 ± 0.010
25%	0.871 ± 0.010	0.858 ± 0.012	0.967 ± 0.003	0.967 ± 0.004
50%	0.908 ± 0.007	0.889 ± 0.005	0.973 ± 0.004	0.965 ± 0.009
95%	0.926 ± 0.005	0.896 ± 0.007	0.978 ± 0.005	0.973 ± 0.003

Does it ~~work~~ *leak?*

Not very much with MEDDOCAN

% for d.a.	MEDDOCAN	
	All 5-grams	Sensitive 5-grams
5%	0.005 ± 0.000	0.008 ± 0.001
10%	0.003 ± 0.000	0.006 ± 0.001
25%	0.003 ± 0.000	0.004 ± 0.000
50%	0.002 ± 0.000	0.003 ± 0.000
95%	0.002 ± 0.000	0.003 ± 0.000
0%	0.001 ± 0.000	0.001 ± 0.000

Does it ~~work~~ *leak?*

Seems like a higher risk for SEPR PHI

% for d.a.	SEPR PHI	
	All 5-grams	Sensitive 5-grams
5%	0.328 ± 0.041	0.233 ± 0.066
10%	0.216 ± 0.002	0.154 ± 0.016
25%	0.183 ± 0.015	0.169 ± 0.021
50%	0.134 ± 0.021	0.141 ± 0.017
95%	0.122 ± 0.013	0.132 ± 0.010
0%	0.028 ± 0.002	0.047 ± 0.002

Analysis

- Models trained using synthetic data performed well!
  - *Slightly* worse – but is it noticeable in practice?
- Data can be synthesized cheaply
  - The synthesizing model can be “small”
  - You need a good model for machine annotation
- But: the synthetic data contained memorized n-grams
  - Crude metric – but gives some indication

46

Kiefer et al. (RANLP 2025)

Below is an instruction that describes a task, paired with an input that provides further context. Write a response that appropriately completes the request.

Instruction:

Given a list of textual descriptions of procedure and diagnosis codes, generate a corresponding clinical discharge summary that provides comprehensive and relevant details about the patient's medical history, current condition and treatment received at the hospital.

Input:

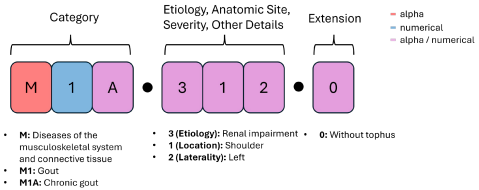
["STEMI, unspecified site", "Primary hypertension", "Unspecified hyperlipidemia", "Coronary artery stent placement, drug-eluting", "Cardiac output monitoring", "IV thrombolytic administration"]

Response:

Discharge Summary

Discharge Summary

ICD-10 Codes

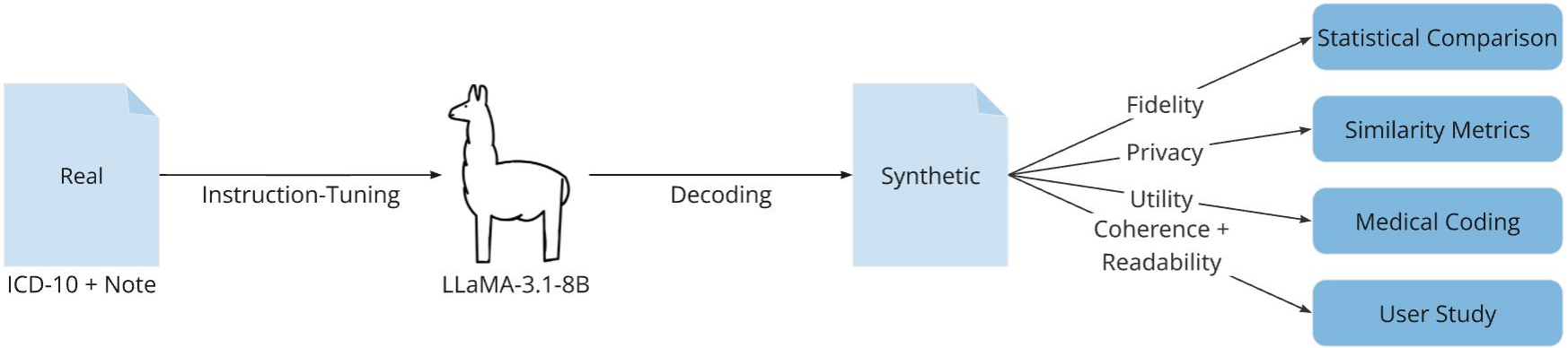


Chronic gout due to renal impairment, left shoulder, without tophus

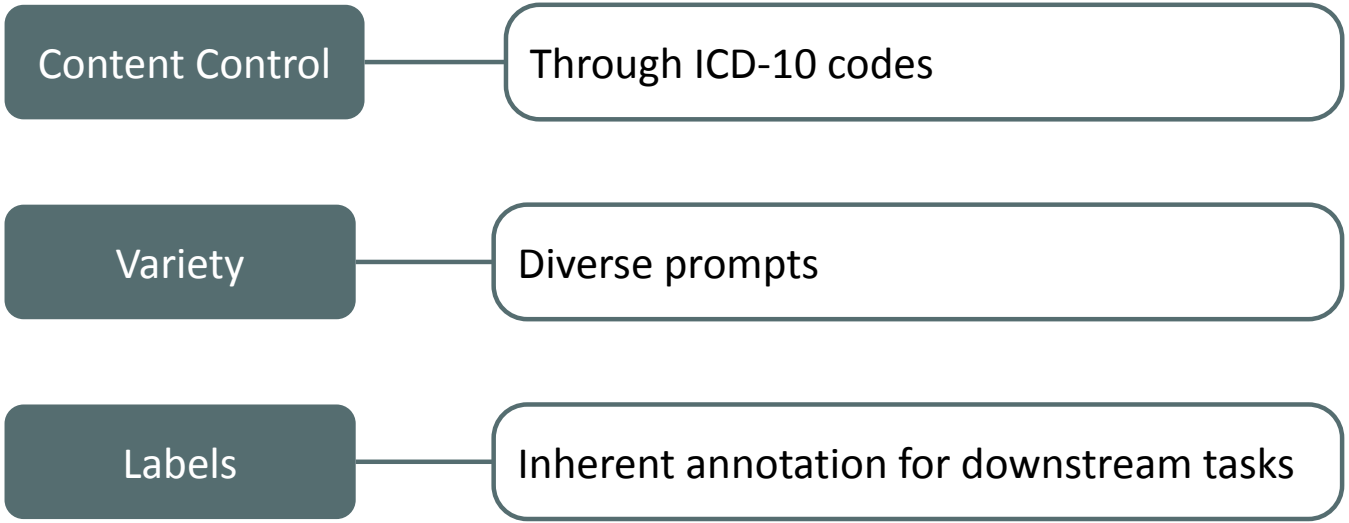
47

Kiefer et al. (RANLP 2025)

Methodological overview



Why ICD-10 codes?

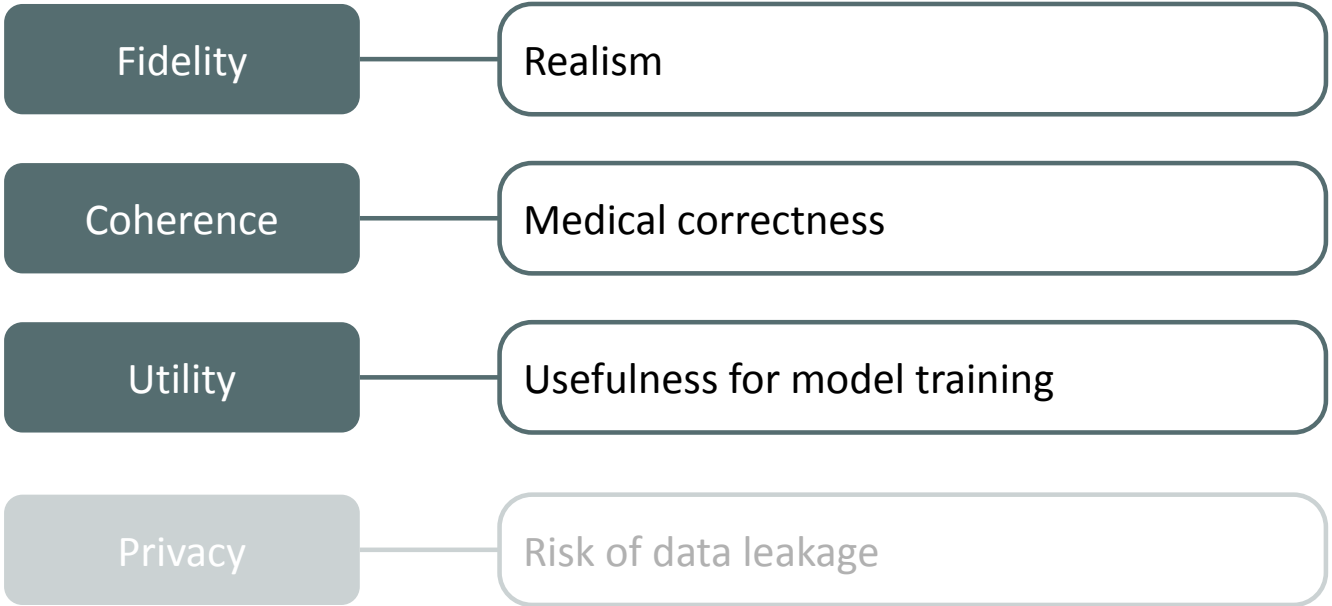


Create synthetic datasets

- Synthetic MIMIC
  - Small set: 13,378
  - Large set: 89,098
- SEPR ICD-10
  - Small set: 47,783
  - Large set: 237,968

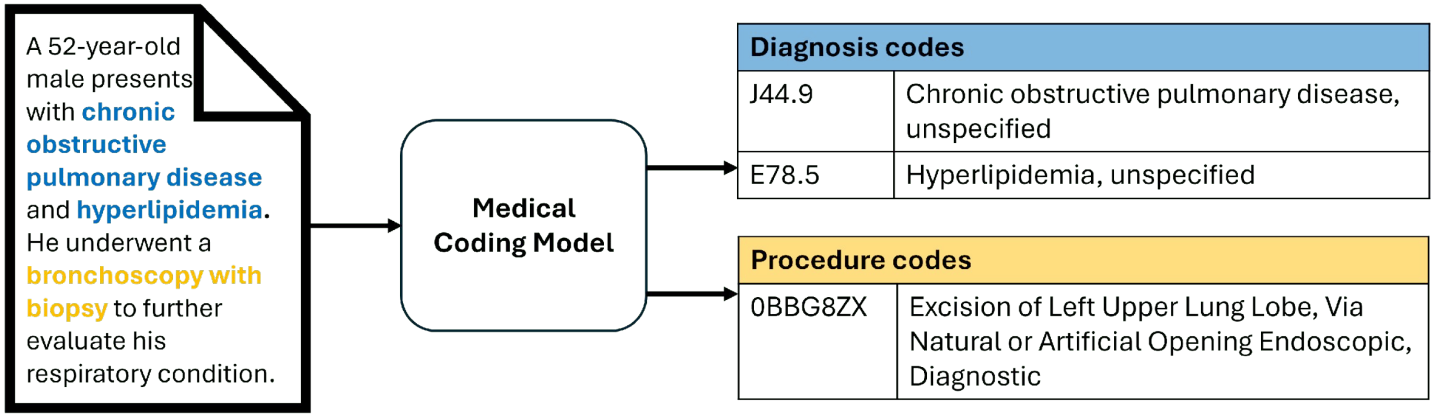
50
Thomas Vakili - thomas.vakili@dsv.su.se

Evaluated dimensions



51

Measuring utility – training a model



52
Thomas Vakili - thomas.vakili@dsv.su.se

Results for utility

Training Data	Classification AUC-ROC Micro	AUC-ROC Macro	F_1 Micro
Edin et al. (2023)	99.2±0.0	96.6±0.2	58.5±0.7
Real MIMIC-L Short	99.2±0.0	96.6±0.1	58.7±0.4
Synth MIMIC-L	**98.9±0.0	**94.8±0.0	*54.8±0.8
Real MIMIC-S Short	96.9±0.4	83.5±1.7	48.2±1.7
Synth MIMIC-S	95.4±0.2	*77.0±0.9	*37.6±0.6

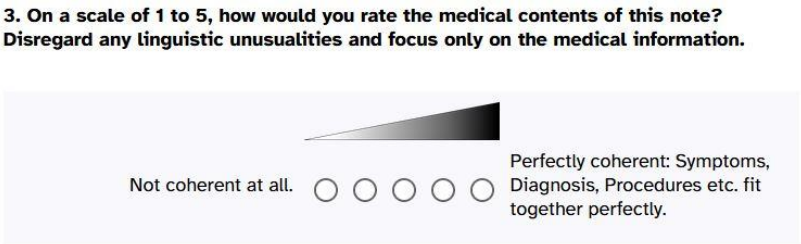
Training Data	Classification AUC-ROC Micro	AUC-ROC Macro	F1 Micro
Lamproudis et al. (2024)	-	-	61.0
Real SEPR-L	99.3±0.1	97.0±0.2	60.2±0.9
Synth SEPR-L	98.8±0.2	**95.3±0.1	**54.7±0.6
Real SEPR-M	98.5±0.0	92.2±1.2	52.4±0.5
Synth SEPR-M	*98.1±0.1	89.7±0.8	**45.9±1.1

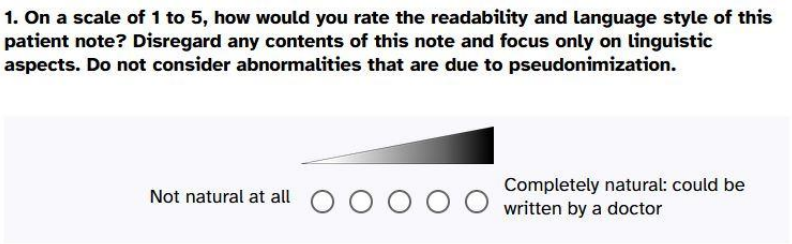
53

Thomas Vakili - thomas.vakili@dsv.su.se

User study: fidelity and coherence

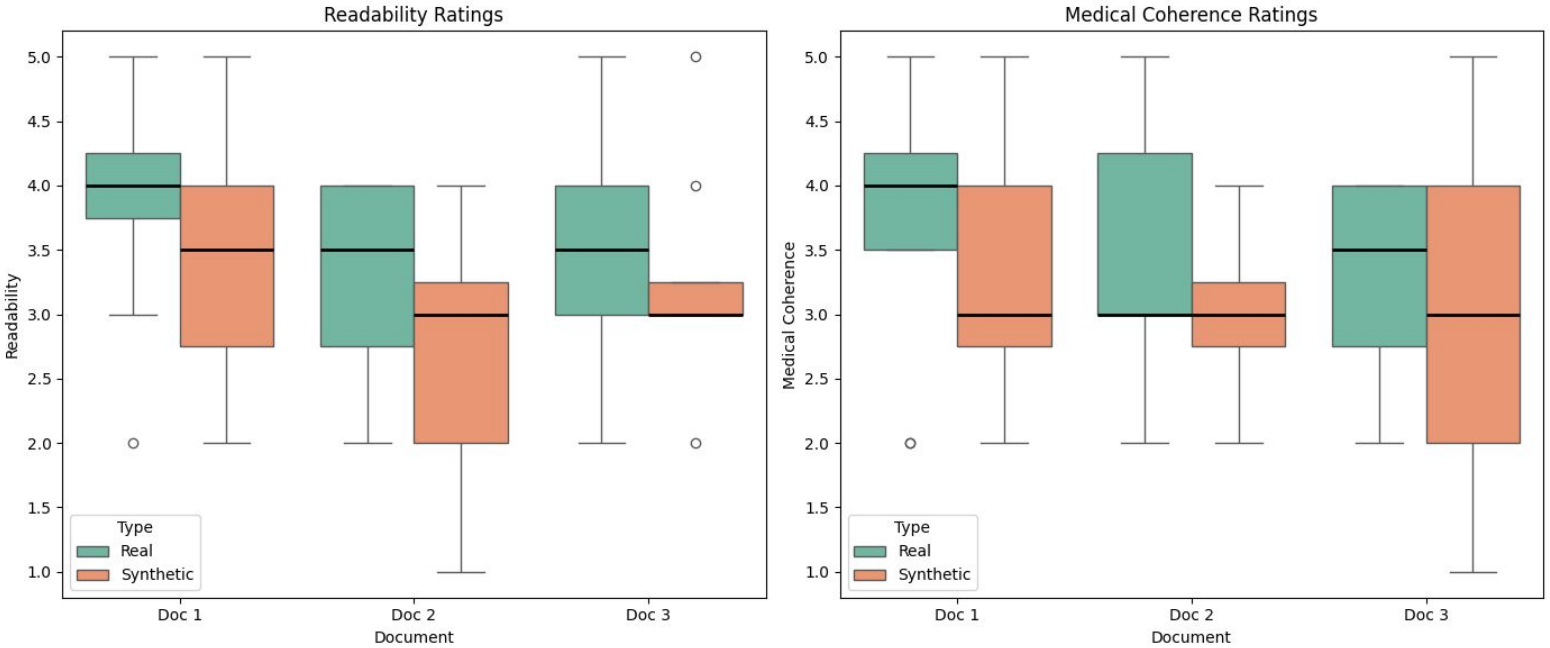
- **Participants:** Medical experts (doctors and nurses)
- **Samples:** Real and synthetic documents belonging to the same ICD-10 codes
- **Task 1:** Rating readability
- **Task 2:** Rating medical coherence





54

Results from the user study



55

Analysis

- Synthetic data resulted in worse (but competitive) models
  - Additional experiments attribute this to noise
- Lower coherence and readability
  - Low sample size and inter-annotator agreement
  - Not statistically significant, but reasonable

Thomas Vakili - thomas.vakili@dsv.su.se

Stockholm University

Conclusions

Reflections on synthesis

- Models trained using synthetic data performed well!
  - *Slightly* worse – but managable (most of the time)
- Data can be synthesized cheaply
  - The synthesizing model can be “small”
  - You need a good model for machine annotation
- But: difficult to know what the privacy gains are

Future work

- Do the results hold for harder clinical problems?
- Does synthesis lead to more bias in the EHRs?
- Do synthetic records preserve clinical information?
  - *Want to help us out? Contact me!*

Thomas Vakili - thomas.vakili@dsv.su.se

Stockholm University

Thank you!