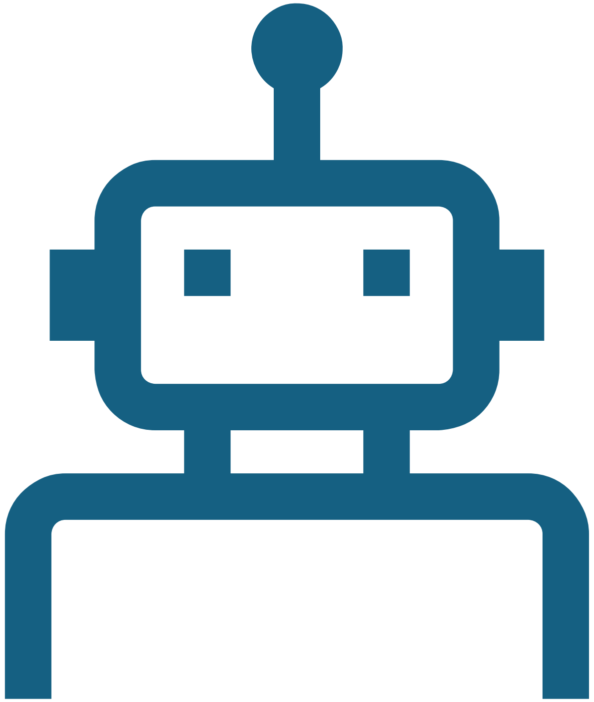


Chatbot Training for Healthcare Applications

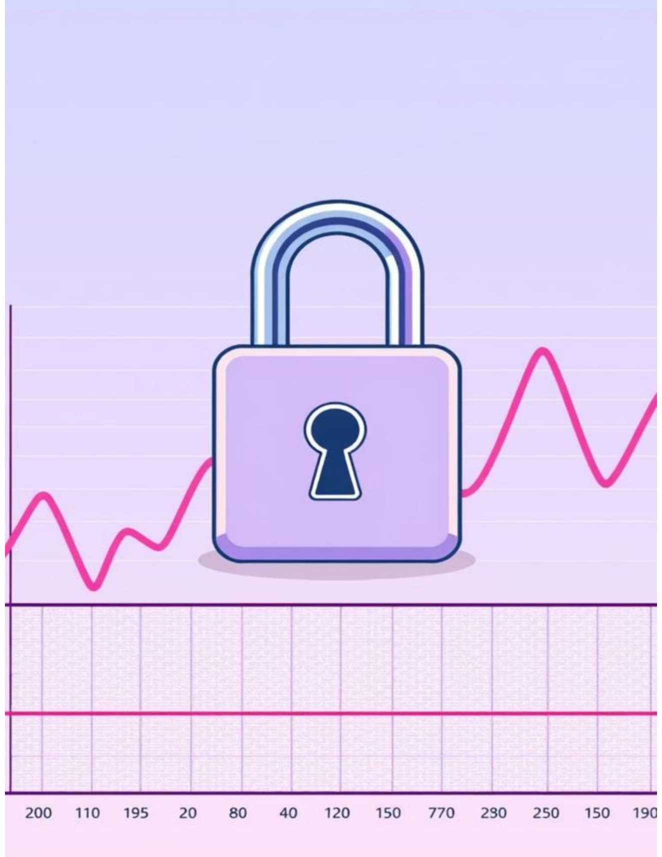
PRIVACY & ANONYMIZATION

- *Suryalaxmi Ravianandan*

Agenda

01	Potential Privacy Risks
02	GDPR Compliance
03	Anonymization Techniques

Why Data Governance Matters in Healthcare AI



Transformative Potential

AI chatbots like ChatGPT can revolutionise patient care but rely on vast amounts of sensitive data.

Uniquely Sensitive Data

Medical data, including biometric and health information, cannot be changed if leaked, making it extremely vulnerable.

Historical Vulnerabilities

Healthcare data breaches affected over 249 million patients between 2005-2019, highlighting systemic weaknesses.

https://pmc.ncbi.nlm.nih.gov/articles/PMC7349636/

Regulatory Imperative

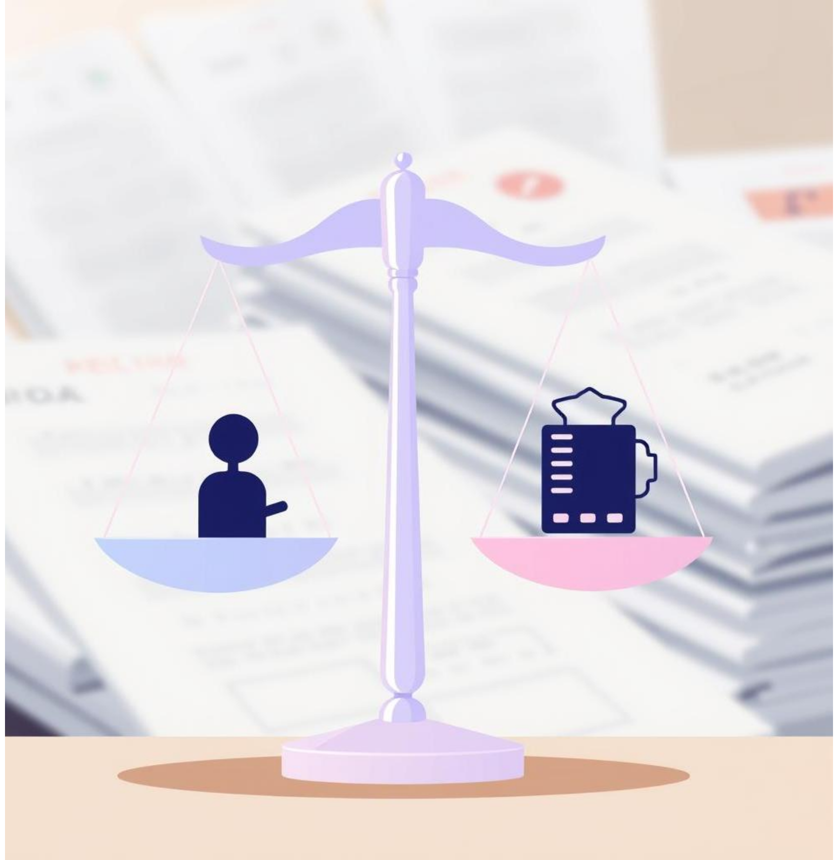
Strong governance is essential to protect patient trust and comply with stringent regulations like HIPAA and GDPR.

Potential Privacy Risks

- Sensitive Data Exposure
- Unauthorized Access
- Data Breaches
- Re-identification
- Improper Data Retention

Regulatory & Ethical Compliance Essentials

- ✓ HIPAA and GDPR mandate strict controls on PHI use and sharing within AI applications.
- ✓ Healthcare providers must ensure chatbot inputs are fully de-identified before processing.
- ✓ Continuous monitoring and regular risk assessments are critical to maintaining compliance and adapting to evolving threats.
- ✓ Transparency with patients about AI use builds trust and supports the delivery of ethical and responsible care.



Anonymization Techniques

Pseudonymization	Tokenization	Redaction	Differential Privacy
Replace direct identifiers with pseudonyms or codes.	Replace identifiers with placeholder tokens.	Remove sensitive data	Inject noise into the dataset
Limitation: not fully anonymous.	Limitation: Requires secure token vault	Limitation: Reduces context for training	Limitation: Complex to implement, may harm accuracy

Reversibility:

- ✓ Pseudo/Token = reversible with key;
- ✓ Redaction/DP = irreversible.

Anonymization Techniques

Pseudonymization	Tokenization	Redaction	Differential Privacy
Replace direct identifiers with pseudonyms or codes.	Replace identifiers with placeholder tokens.	Remove sensitive data	Inject noise into the dataset
Limitation: not fully anonymous.	Limitation: Requires secure token vault	Limitation: Reduces context for training	Limitation: Complex to implement, may harm accuracy

Workflow:

Ingestion → Pseudo/Token/Redact → k-/l-checks → DP for outputs.

Dataset : Medical Student Mental Health

```python
# Load data
data_path = "/content/drive/MyDrive/Data.csv"
df = pd.read_csv(data_path)

df.head()
```

	id	age	year	sex	glang	part	job	stud_h	health	psyt	jspe	qcae_cog	qcae_aff	amsp	erec_mean	cesd	stai_t	mbi_ex	mbi_cy	mbi_ea
0	2	18	1	1	120	1	0	56	3	0	88	62	27	17	0.738095	34	61	17	13	20
1	4	26	4	1	1	1	0	20	4	0	109	55	37	22	0.690476	7	33	14	11	26
2	9	21	3	2	1	0	0	36	3	0	106	64	39	17	0.690476	25	73	24	7	23
3	10	21	2	2	1	0	1	51	5	0	101	52	33	18	0.833333	17	48	16	10	21
4	13	21	3	1	1	1	0	22	4	0	102	58	28	21	0.690476	14	46	22	14	23

Variable Name	Variable Label	Variable Scale
id	Participants ID number	string
age	age at questionnaire 20-21	numeric
year	CURICULUM YEAR : In which curriculum year are you?	1=Bmed1; 2=Bmed2; 3=
sex	GENDER : To which gender do you identify the most ?	1=Man; 2=Woman; 3=N
glang	MOTHER TONGUE: What is your mother tongue?	1=French; 15=German;
part	PARTNERSHIP STATUS : Do you have a partner?	0=No; 1=Yes
job	HAVING A JOB : Do you have a paid job?	0=No; 1=Yes
stud_h	HOURS OF STUDY PER WEEK : On average, how many hours per week do you study on top of courses?	
health	SATISFACTION WITH HEALTH : How satisfied are you with your health?	1=Verydissatisfied; 2=Di
psyt	PSYCHOTHERAPY LAST YEAR : During the last 12 months, have you ever consulted a psychotherapist or a ps	0=No; 1=Yes
jspe	JSPE total empathy score	numeric
qcae_cog	QCAE Cognitive empathy score	numeric
qcae_aff	QCAE Affective empathy score	numeric
amsp	AMSP total score	numeric
erec_mean	GERT : mean value of correct responses	numeric
cesd	CES-D total score	numeric
stai_t	STAI score	numeric
mbi_ex	MBI Emotional Exhaustion	numeric
mbi_cy	MBI Cynicism	numeric
mbi_ea	MBI Academic Efficacy	numeric

Anonymization Techniques

1. Tokenization - process of replacing sensitive data elements with Tokens

✓ In data privacy, tokenization is used to protect information such as names, IDs, or other identifiers in a dataset.

- "John Smith", "123-45-6789", or a unique user ID
- **Token are** randomly generated or mapped value (e.g., "A1B2C3D4", "abc123", or "user_001") that replaces the original data in the dataset.

✓ **Mapping :**

- The relationship between tokens and real values is securely stored in a separate, protected location (the "token vault").
- Without access to this vault, the token cannot be reversed to reveal the original value.

➢ **Protects Sensitive Data:** Even if the dataset is exposed, the original values are not revealed.

➢ **Prevents Re-identification:** Tokens cannot be reversed without the mapping vault.

➢ **Compliance:** Helps meet privacy regulations by reducing the risk of data breaches.

Anonymization Techniques - Tokenization

index	id	age	year	sex	glang	part	job	stud_h	health	psyt	jspe	qcae_cog	qcae_aff	amsp	erec_mean	cesd	stai_t	mbi_ex	mbi_cy	mbi_ea
0	2	18	1	1	120	1	0	56	3	0	88	62	27	17	0.73809522	34	61	17	13	20
1	4	26	4	1	1	1	0	20	4	0	109	55	37	22	0.69047618	7	33	14	11	26
2	9	21	3	2	1	0	0	36	3	0	106	64	39	17	0.69047618	25	73	24	7	23
3	10	21	2	2	1	0	1	51	5	0	101	52	33	18	0.83333331	17	48	16	10	21
4	13	21	3	1	1	1	0	22	4	0	102	58	28	21	0.69047618	14	46	22	14	23

index	age	year	sex	glang	part	job	stud_h	health	psyt	jspe	qcae_cog	qcae_aff	amsp	erec_mean	cesd	stai_t	mbi_ex	mbi_cy	mbi_ea	id_token
0	18	1	1	120	1	0	56	3	0	88	62	27	17	0.73809522	34	61	17	13	20	zR8nfwgvRU6fPv882CG23A
1	26	4	1	1	1	0	20	4	0	109	55	37	22	0.69047618	7	33	14	11	26	iCuGZlEIRhigfhPzHSyPTA
2	21	3	2	1	0	0	36	3	0	106	64	39	17	0.69047618	25	73	24	7	23	5b310yp6QiWgO88jZ6FcuQ
3	21	2	2	1	0	1	51	5	0	101	52	33	18	0.83333331	17	48	16	10	21	aQxegD0sQGysJ8VU8PVxPA
4	21	3	1	1	1	0	22	4	0	102	58	28	21	0.69047618	14	46	22	14	23	KDZI-SCORA-VPVNttGbaFw

```
id_to_token: [(2, 'zR8nfwgvRU6fPv882CG23A'), (4, 'iCuGZlEIRhigfhPzHSyPTA'), (9, '5b310yp6QiWgO88jZ6FcuQ'), (10, 'aQxegD0sQGysJ8VU8PVxPA'),
token_to_id: [('zR8nfwgvRU6fPv882CG23A', 2), ('iCuGZlEIRhigfhPzHSyPTA', 4), ('5b310yp6QiWgO88jZ6FcuQ', 9), ('aQxegD0sQGysJ8VU8PVxPA', 10),
```

Anonymization Techniques

2. Redaction - process of removing or masking sensitive information from a dataset before it is shared.

✓ **Removing Columns:** Deleting columns that contain direct identifiers

- names, ID numbers, or email addresses.

✓ **Masking Values:** Replacing specific data values with a placeholder

- Replace rare 'lang' values with "REDACTED"

✓ **Partial Redaction:** Hiding only part of a value

- 05.05.1989 as “1989”

➢ **Protect Privacy:** Prevents unauthorized access to personal or identifying information.
➢ **Data Sharing:** Enables sharing of data for research or analysis without exposing confidential details.
➢ **Legal & Ethical Compliance:** Meets requirements of data protection laws and ethical standards.

Anonymization Techniques - Redaction

index	id	age	year	sex	glang	part	job	stud_h	health	psyt	jspe	qcae_cog	qcae_aff	amsp	erec_mean	cesd	stai_t	mbi_ex	mbi_cy	mbi_ea
0	2	18	1	1	120	1	0	56	3	0	88	62	27	17	0.73809522	34	61	17	13	20
1	4	26	4	1	1	1	0	20	4	0	109	55	37	22	0.69047618	7	33	14	11	26
2	9	21	3	2	1	0	0	36	3	0	106	64	39	17	0.69047618	25	73	24	7	23
3	10	21	2	2	1	0	1	51	5	0	101	52	33	18	0.83333331	17	48	16	10	21
4	13	21	3	1	1	1	0	22	4	0	102	58	28	21	0.69047618	14	46	22	14	23

stud_h	health	psyt	jspe	...	erec_mean	cesd	stai_t	mbi_ex	mbi_cy	mbi_ea	id_token	age_group	glang_gen	year_group
56	3	0	88	...	0.738095	34	61	17	13	20	JsiJ_TqeTpiuyJdvx8gMxQ	17-20	Other	Bmed
20	4	0	109	...	0.690476	7	33	14	11	26	GjiB6PtfRGGY5K6uifj-DA	25-29	1	Mmed
36	3	0	106	...	0.690476	25	73	24	7	23	cnl3vFNJSKGb_3Ks9bx76g	21-24	1	Bmed
51	5	0	101	...	0.833333	17	48	16	10	21	bqPJU3ezQ9O9MfhIGJC4Ow	21-24	1	Bmed
22	4	0	102	...	0.690476	14	46	22	14	23	PIAn00-_SDiQRtWOtSz42A	21-24	1	Bmed

Anonymization Techniques

Quasi-Identifiers (QIs) are attributes that, while not directly identifying, can be combined to uniquely pinpoint an individual within a dataset.

✓Defining QIs from our student mental health dataset.

- Age
- curriculum year
- Sex
- mother tongue

✓ Re-identification Risk

A unique combination (e.g., 24-year-old, Non-binary, Mmed, Turkish speaker) can single out a student.

✓ External Linking

This unique combination, when cross-referenced with external data like a class list, compromises privacy.

Anonymization Techniques

k-Anonymity: Ensuring Group Privacy

Every combination of Quasi-Identifiers (QIs) in a dataset must appear in at least k records.

✓ **Violation Example (k=5)**

If a QI combination like (age_group=25–29, year_group=Mmed, sex=Non-binary, glang_gen=Other) appears in only 2 records, it violates 5-anonymity, as 2 is less than 5.

✓ **Strategic Generalisation**

To achieve k-anonymity, QIs must be generalised. This involves broadening categories (e.g., merging age groups 21–24 and 25–29 into 21–29) or removing less critical QIs from enforcement.

✓ **Compliance Achieved**

After generalisation (e.g., dropping 'sex' from enforcement), if the combined group (age_group=21–29, year_group=Mmed, glang_gen=Other) now includes 11 records, 5-anonymity is satisfied (11 ≥ 5).

Anonymization Techniques

l-Diversity

- Within every k-anonymous QI group, the sensitive attribute must have at least l distinct values to prevent re-identification.

✓ **Violation**

QI group (age_group=21–24, year_group=Bmed, sex=Woman, glang_gen=French) with 7 records.

If all 7 have 'psyt=0', l-diversity with l=2 fails

✓ **How to fix:**

a. **Generalise QIs** - Broaden categories to mix records and increase diversity.

b. **Adjust Sensitive Attribute** - Choose a sensitive attribute with more variation or coarsen it.

Implementation - Anonymization Techniques

```
 before– k-anonymity: 155 groups
Before – k-anonymity (head):
age_group  year_group  sex         glang_gen
17–20      Bmed        Non–binary  1            0
                                   90           0
                       Man         20           0
           Mmed        Man         1            0
                                   Other        0
           Bmed        Non–binary  20           0
                                   15           0
                                   102          0
           Mmed        Man         102          0
                                   90           0
dtype: int64
before – l-diversity: 138 groups

Before – l-diversity (head):
age_group  year_group  sex         glang_gen
17–20      Bmed        Non–binary  1            0
                                   90           0
                       Man         20           0
           Mmed        Man         1            0
                                   Other        0
           Bmed        Non–binary  20           0
                                   15           0
                                   102          0
           Mmed        Man         102          0
                                   90           0
Name: health, dtype: int64
```

```
After – k-anonymisation: 352 groups
age_group  year_group  sex    glang_gen
40+        Bmed        Woman  Other       0
                              20          0
                              15          0
                              102         0
                              90          0
                              1           0
                              NaN         0
                       Man    Other       0
                              20          0
                              15          0
dtype: int64
After – l-diversity: 352 groups
age_group  year_group  sex    glang_gen
40+        Bmed        Woman  Other       0
                              20          0
                              15          0
                              102         0
                              90          0
                              1           0
                              NaN         0
                       Man    Other       0
                              20          0
                              15          0
Name: health, dtype: int64
Suppressed rows (any QI NaN): 8.92%
```

Advanced Anonymization with LLMs

- Recent LLMs, such as Llama-3 70B, have demonstrated remarkable capabilities, achieving a 99.24% success rate in automatically removing PHI from clinical text. This breakthrough is pivotal for secure healthcare AI development.

- Automating anonymisation enables the scalable and secure use of vast amounts of unstructured medical data, accelerating research and development without compromising patient privacy.

Deidentifying Medical Documents with Local, Privacy-Preserving Large Language Models: The LLM-Anonymizer

Isabella C. Wiest, M.D., M.Sc.,[1,2] Marie-Elisabeth Leßmann, M.D.,[1,3] Fabian Wolf, M.Sc.,[1] Dyke Ferber, M.D.,[1,4] Marko Van Treeck, M.Sc.,[1] Jiefu Zhu, M.Sc.,[1] Matthias P. Ebert, M.D.,[2,5,6] Christoph Benedikt Westphalen, M.D.,[7,8,9] Martin Wermke, M.D.,[3] and Jakob Nikolas Kather, M.D., M.Sc.[1,3,4]

Received: May 26, 2024; Revised: November 18, 2024; Accepted: December 29, 2024; Published: March 27, 2025

Abstract

BACKGROUND Medical research with real-world clinical data is challenging as a result of privacy requirements. Patient data should be anonymized before analysis in research studies. Anonymization procedures aim to reduce the reidentification risk below a certain threshold, while maintaining the usefulness of the data for research purposes. However, in the context of medical text, these procedures are notoriously hard to automate and, therefore, are not scalable. Recent advancements in natural language processing (NLP), driven by the development of large language models (LLMs), have markedly improved the automatic processing of unstructured text.

METHODS We hypothesize that LLMs are highly effective tools for extracting patient-related information, which can subsequently be used to remove personal information from medical reports, while at the same time preserving information required for downstream research purposes. To test this, we conducted a benchmark study using eight local LLMs (Llama-3 8B, Llama-3 70B, Llama-2 7B, Llama-2 70B, Llama-2 7B Sauerkraut, Llama-2 70B Sauerkraut, Mistral 7B, and Phi-3 Mini) to extract and remove patient-related information from a dataset of 250 real-world clinical letters.

RESULTS Our results demonstrate that our LLM-Anonymizer, when used with Llama-3 70B, achieved a success rate of 99.24% in removing text characters carrying personal identifying information. It missed only 0.76% of text characters with identified personal information and mistakenly redacted 2.43% of characters.

CONCLUSION We provide our full LLM-based Anonymizer pipeline under an open-source

Drs. Wiest and Leßmann contributed equally to this article.

The author affiliations are listed at the end of the article.

Conclusion

Safe AI in Healthcare is Possible with Rigorous Data Governance

- **Privacy and anonymisation** remain essential for healthcare chatbots, safeguarding patients and guaranteeing ethical handling of data.

- **Utilising advanced anonymisation techniques and privacy-enhancing technologies (PETs)** allows for innovation while managing risk, promoting secure progress in technology.

- The **Medical Student Mental Health dataset** serves as a prime example of responsible AI training data usage, establishing a standard for upcoming initiatives.

It is important to prioritise patient data protection while embracing AI's potential to revolutionise healthcare delivery globally.