Friedrich-Alexander-Universität
Technische Fakultät

M.Sc. In Data Science
Data Governance Applications

Chatbot Training for Healthcare Applications

Agenda

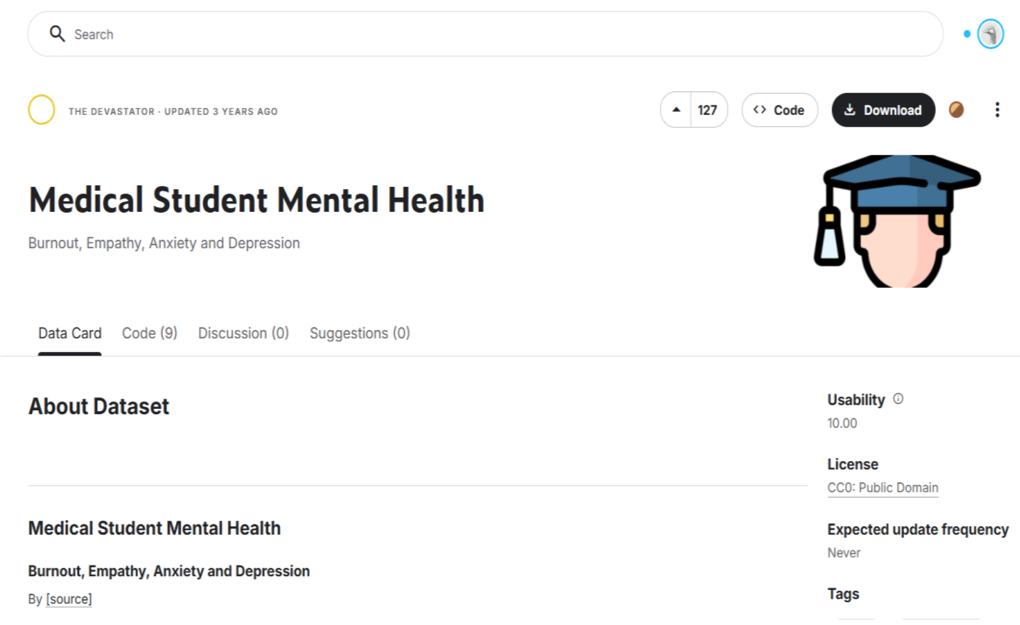
Introduction

- A new AI-powered chatbot is being developed to assist patients with healthcare inquiries.

- **Goal:** The chatbot aims to improve patient access to information, help with appointment scheduling, and enhance engagement.

- **The Mandate:** Our team has been brought in as a consulting group to conduct a thorough audit before the system is deployed.

Data Documentation



https://www.kaggle.com/datasets/thedevastator/medical-student-mental-health/data

Author : "thedevastator"

Purpose :

- Survey data about mental health among medical students.
- Useful for analyzing prevalence of mental health issues (depression, anxiety, panic attacks), associations with demographic, academic, and behavioral variables;
- Possibly modeling whether students seek professional mental health treatment.

Size of Dataset	55 MB
Number of Instances	(888 x 20)
Number of Fields	888
Labeled Classes	20
Number of Labels	20

Data Documentation

id	Participants ID number	string
age	age at questionnaire 20-21	numeric
year	CURICULUM YEAR : In which curriculum year are...	1=Bmed1; 2=Bmed2; 3=Bmed3; 4=Mmed1; 5=Mmed2; 6...
sex	GENDER : To which gender do you identify the most?	1=Man; 2=Woman; 3=Non-binary
glang	MOTHER TONGUE: What is your mother tongue?	1=French; 15=German; 20=English; 37=Arab; 51=B...
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...	string
health	SATISFACTION WITH HEALTH : How satisfied are you...?	1=Verydissatisfied; 2=Dissatisfied; 3=Neither...

psyt	PSYCHOTHERAPY LAST YEAR : During the last 12 months...	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

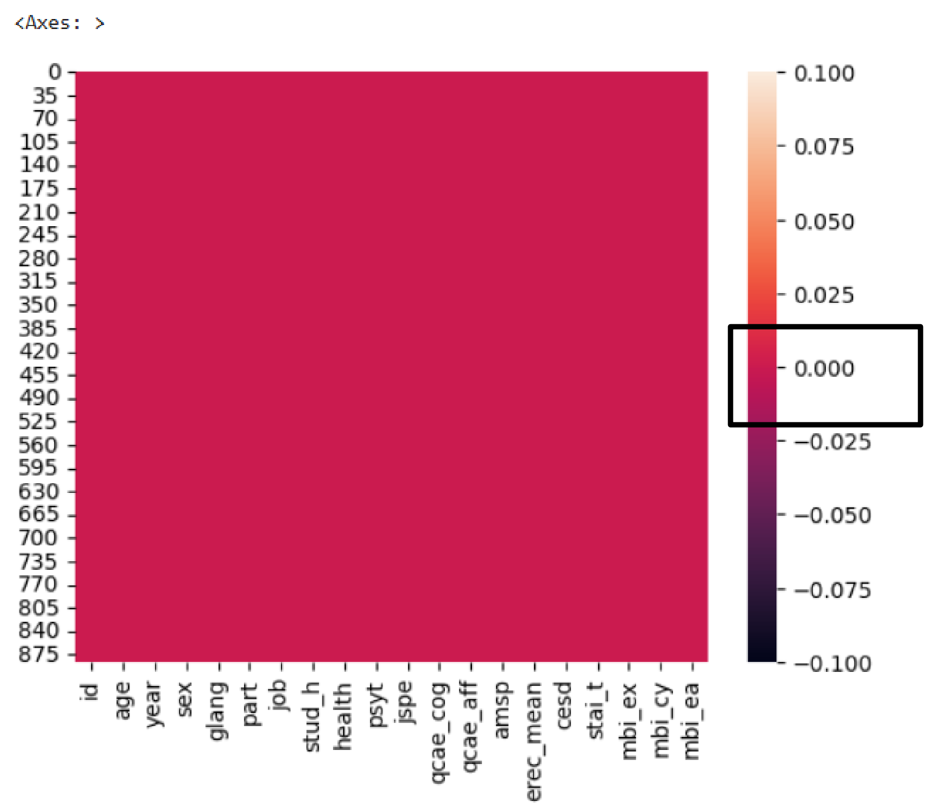
Data Quality Evaluation

Visualizing and Assessment of Missing Values

- The Non-Null Count column in the pandas DataFrame output shows that every single one of the 886 entries has a value.

- This is visually confirmed by the heatmap on the right, which is a solid block of color, indicating the absence of any missing data points.

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 886 entries, 0 to 885
Data columns (total 20 columns):
 #   Column     Non-Null Count  Dtype  
---  ------     --------------  -----  
 0   id         886 non-null    int64  
 1   age        886 non-null    int64  
 2   year       886 non-null    int64  
 3   sex        886 non-null    int64  
 4   glang      886 non-null    int64  
 5   part       886 non-null    int64  
 6   job        886 non-null    int64  
 7   stud_h     886 non-null    int64  
 8   health     886 non-null    int64  
 9   psyt       886 non-null    int64  
 10  jspe       886 non-null    int64  
 11  qcae_cog   886 non-null    int64  
 12  qcae_aff   886 non-null    int64  
 13  amsp       886 non-null    int64  
 14  erec_mean  886 non-null    float64
 15  cesd       886 non-null    int64  
 16  stai_t     886 non-null    int64  
 17  mbi_ex     886 non-null    int64  
 18  mbi_cy     886 non-null    int64  
 19  mbi_ea     886 non-null    int64  
dtypes: float64(1), int64(19)
memory usage: 138.6 KB
```



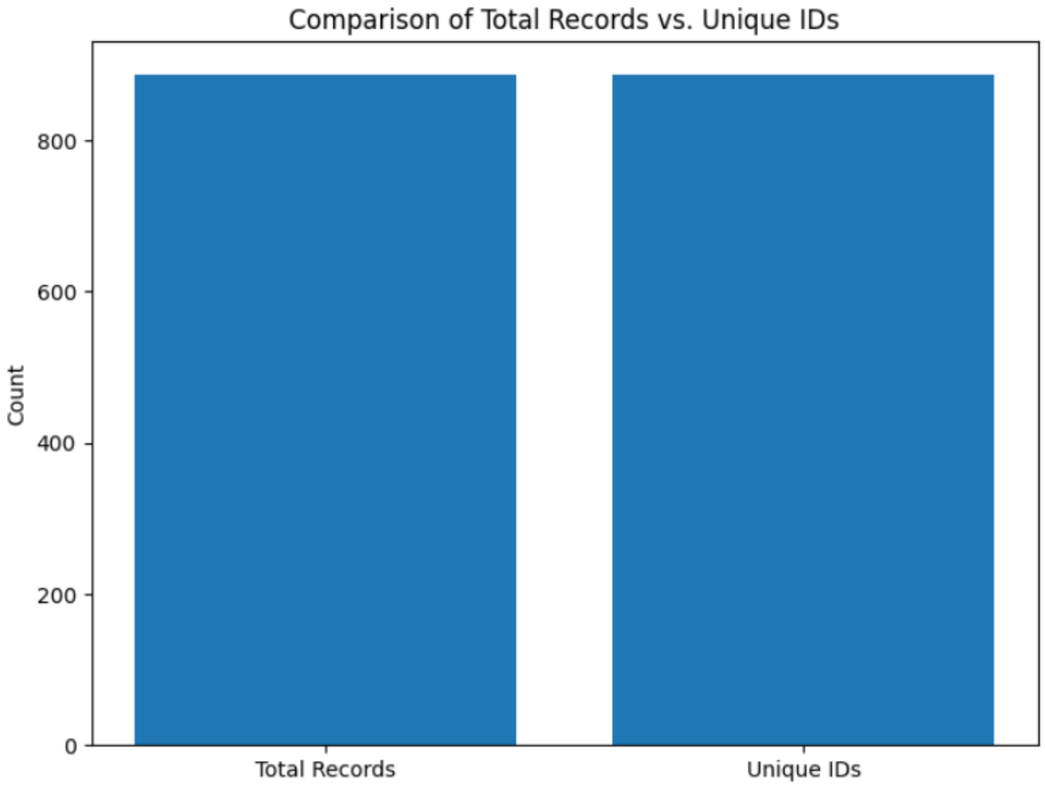
Data Quality Evaluation

Visualizing and Assessment of Duplicate Records

- The bar chart shows the count of Total Records is exactly equal to the count of Unique IDs, proving that every entry is unique.
- Thus confirming that the dataset contains no duplicate records.

```
codebook['id'].duplicated()
```

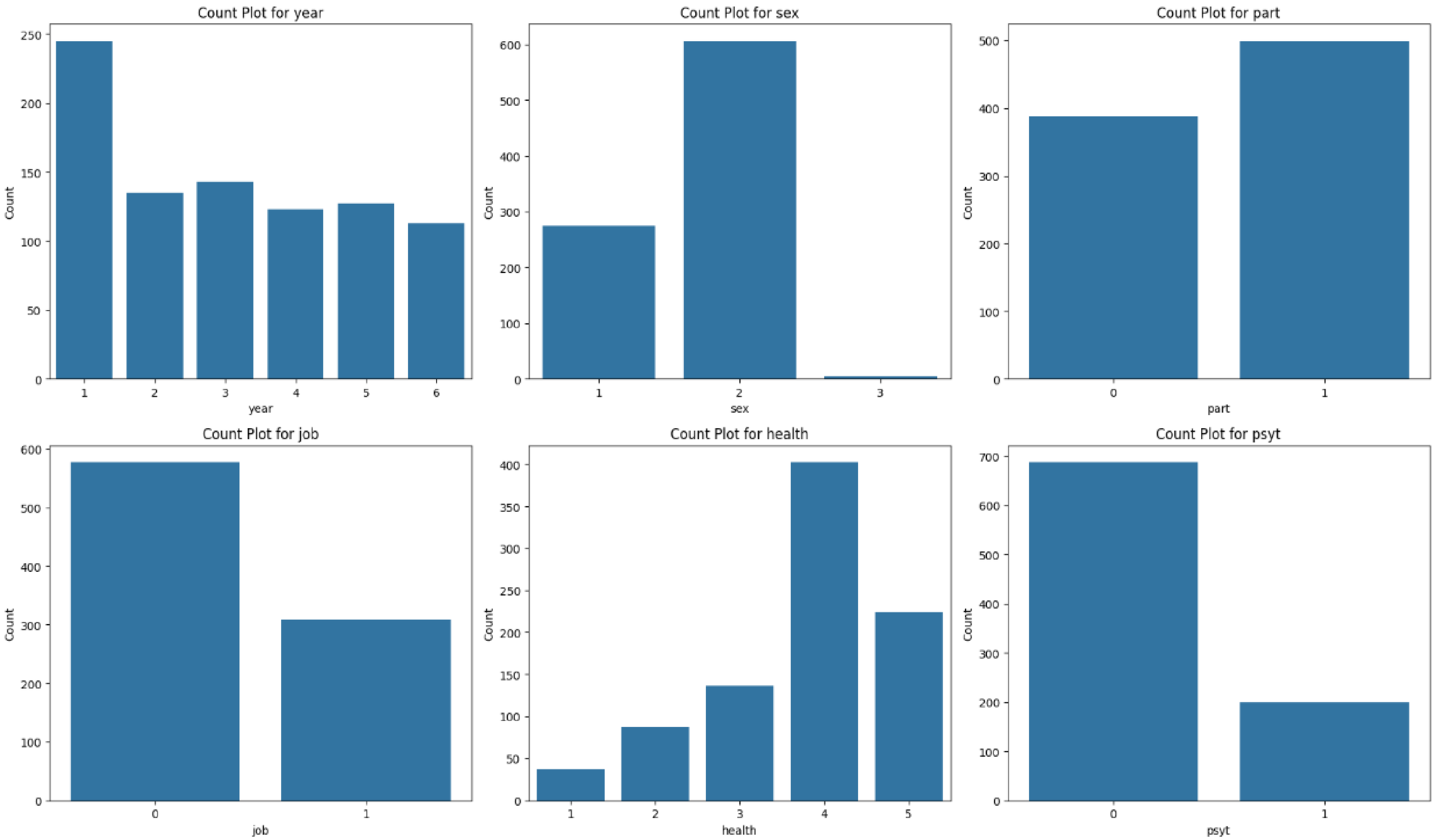
```
0      False
1      False
2      False
3      False
4      False
       ...
881    False
882    False
883    False
884    False
885    False
Name: id, Length: 886, dtype: bool
```



Data Quality Evaluation

Visualizing and Assessment of Categorical Data

The charts reveal the frequency of each category within key variables, highlighting an imbalance in the distribution of students across categories, such as a majority being female and having high health scores.



Potential Privacy Risks

Sensitive Data Exposure

Unauthorized Access

Data Breaches

Re-identification

Improper Data Retention

Anonymization

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

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.

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

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

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

k-Anonymity: Ensuring Group Privacy
Every combination of Quasi-Identifiers (QIs) in a dataset must appear in at least k records.

l-Diversity

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

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

FAU

➢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.,
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 articl

The author affiliations are listed the end of the article.

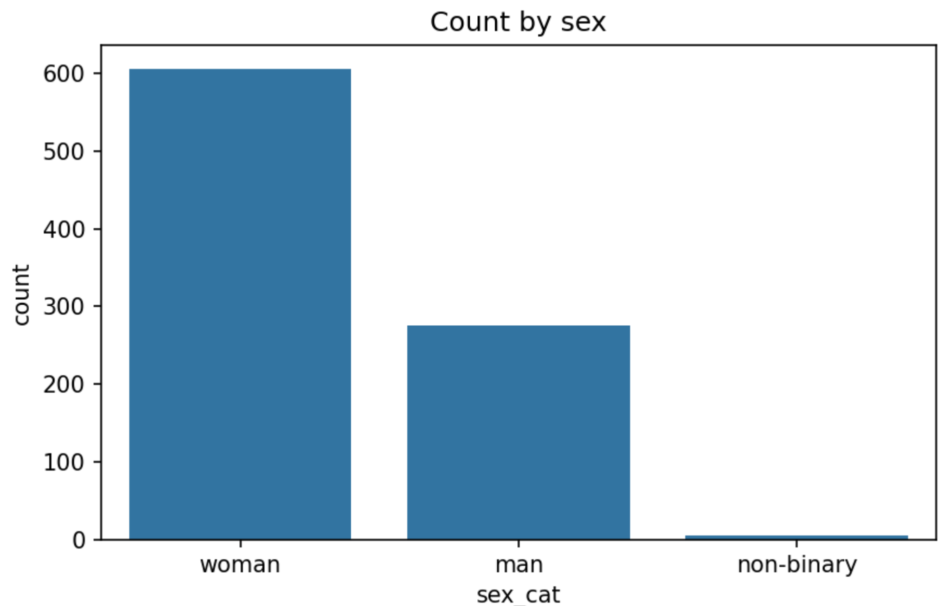
Bias & Fairness Audit — Context

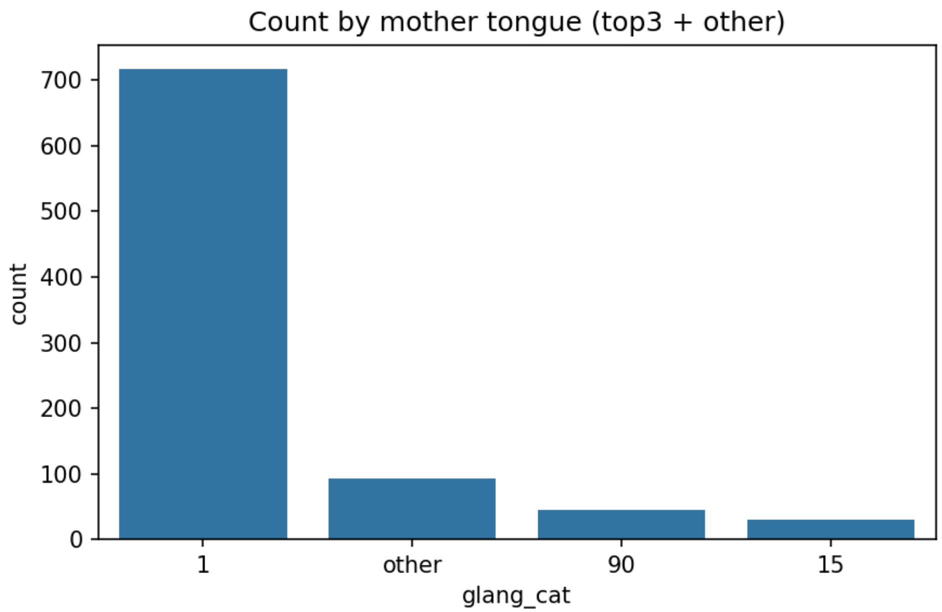
Evidence, Mitigation & Recommendations

- Healthcare chatbots must treat patients fairly → errors = real harm.
- Risk: missed burnout in minority groups → unsafe care.
- Our focus: fairness audit across sex & language, plus a mitigation test

Key evidence: Sample imbalance

- Majority = women + one dominant mother tongue.
- Non-binary + minority languages = tiny groups.
- → Large imbalance → subgroup results unreliable.

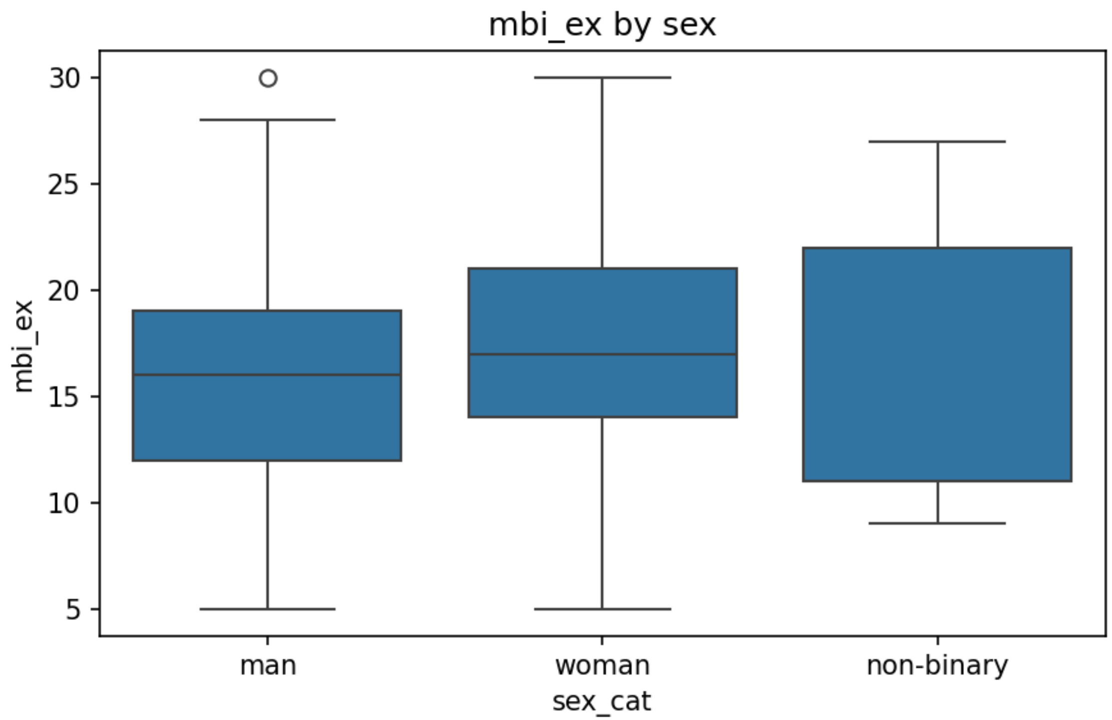




Key Evidence : Outcome disparity

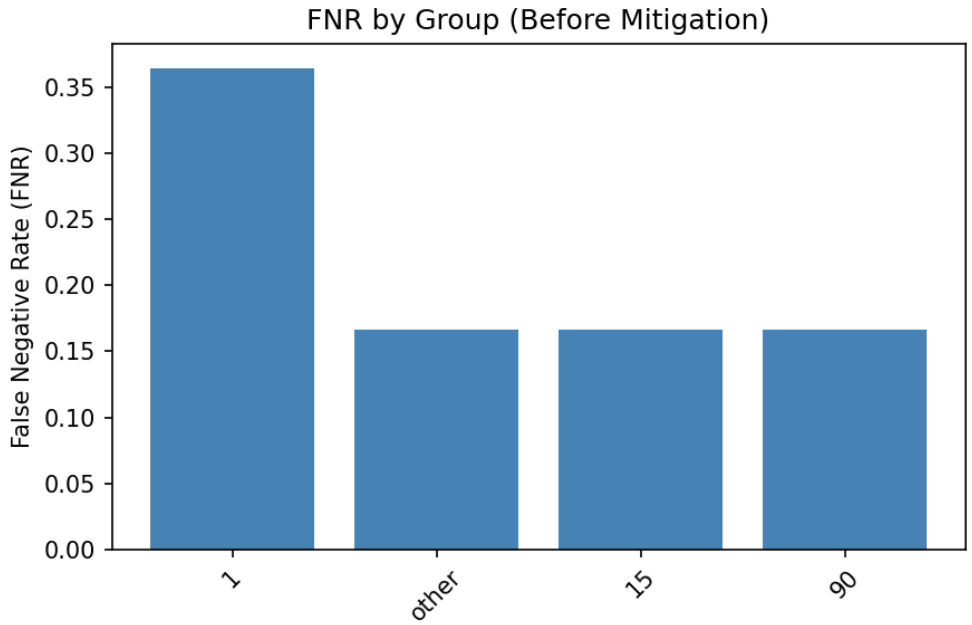
Emotional exhaustion (mbi_ex) by sex

- Burnout (emotional exhaustion) higher in women vs men.
- Non-binary variance very high, but sample is tiny.
- → Subgroups experience burnout differently.



Model fairness: (Before Mitigation)

- Logistic regression with age, year, sex, language.
- Key metric: False Negative Rate (FNR) → missed burnout = unsafe.
- Results:
  - Women: FNR ~36%
  - Minority languages: FNR ~16% (unstable).
- → Model misses high-burnout cases unevenly.

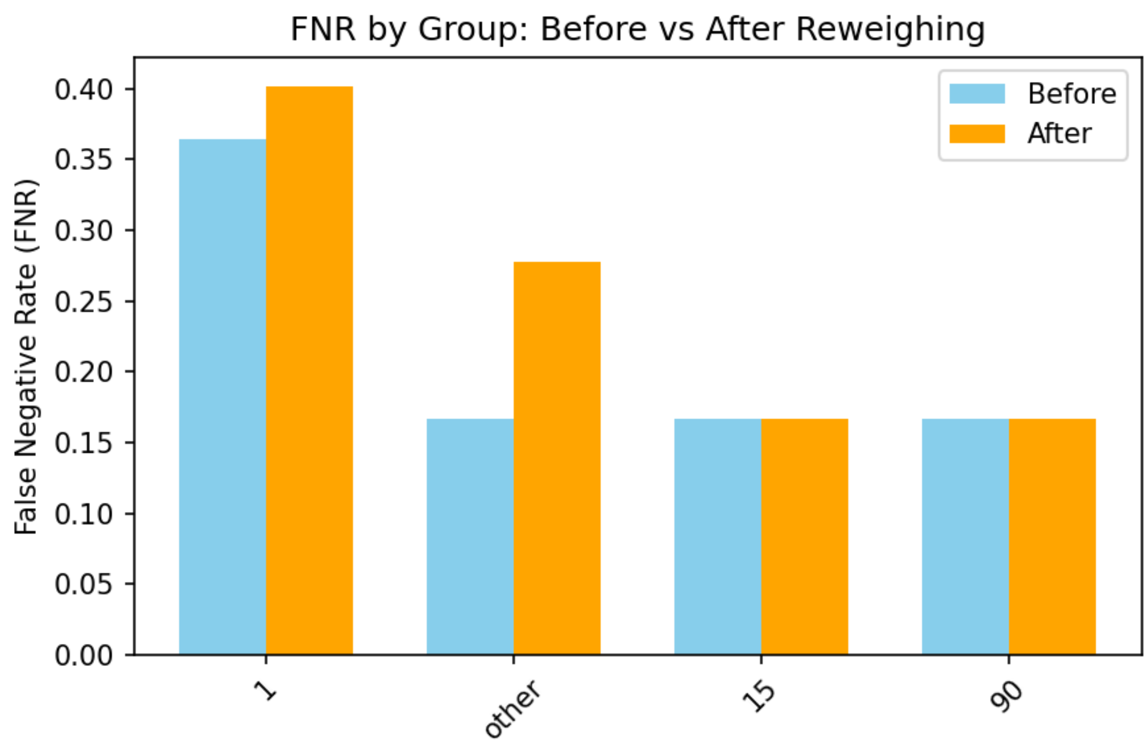


Dominant groups have high false negatives (unsafe).
Minority results look 'better' but are unreliable due to very small sample sizes

Mitigation Experiment — Reweighing

- Reweighing ↑ fairness for minority groups.
- Overall AUC: 0.562 → 0.570 (small change).
- Trade-off: fairness ↑ but accuracy slightly ↓.
- Reweighing reduced disparities but introduced a trade-off:

higher equity for smaller groups
at the cost of a small performance drop.



Trade-offs & Limitations

- Trade-offs:
  - ✅ Fairness ↑ for minority groups.
  - ❌ Accuracy ↓ slightly → more false alarms.
- Stakeholders:
  - Patients → safer, fewer missed burnout cases.
  - Doctors/Nurses → more workload from false alarms.
  - Hospital → balance safety vs efficiency
- Limitations: tiny non-binary group; survey ≠ clinical dataset.

Recommendations

Immediate (pre-rollout):

- Publish per-group fairness metrics.
- Use class weights & conservative thresholds.
- Route high-risk outputs to human review.

Medium-term:

- Collect more subgroup data (non-binary, minority languages).
- Test fairness-aware training approaches.
- Define fairness thresholds with clinicians.

Final Conclusion

Bias exists: dataset heavily imbalanced (sex, language).

Disparities matter: higher burnout in women, underrepresented groups unstable.

Model fairness: uneven FNR → unsafe for deployment "as is."

Mitigation works (partially): reweighing reduced gaps but lowered AUC slightly.

Fairness = trade-off: safety vs efficiency, must be clinician-guided.

Conclusion

Thank you!

Questions?