Supplementary Information

Supplementary Tables

Supplementary Table 1. Annotation examples for each SHI type. The text to be annotated is indicated in bold font.

Main Category	Sub Category	Examples
Name	PATIENT	I had the pleasure to see Mrs. **Kathy** today.
	DOCTOR	(**HJ**/ta 18/3/34) MICROSCOPIC (reported by Dr **F Boa**):
Location	HOSPITAL	to**St. Gollan Hospital**
	DEPARTMENT	**8.MEDIC/SURGERY WARD**- STGH Slides sent to **South Pathology Central**
	ROOM	**Room 12** **OBG floor, Room8**
	STREET	**20 King Avenue** **10 Peach Street**
	CITY	**WickVille** **Pondy**
	STATE	**NSW** **ACT**
	ZIP	**8844** **8867**
	COUNTRY	**Australia** **USA**
	ORGANIZATION	**Oracle** **Microsoft**
	LOCATION-OTHER	**P.O. BOX-132**
Age	AGE	**92**yr old **40** yr old, squamous cervix carcinoma.
ID	IDNUM	Episode No: **21DC43234B** Lab No: **92BAH12861**
	MEDICAL_RECORD	**712213.KMUH** **441774. TMUH**
Contact	PHONE	**02-2788-3785** **8292 0000**
	URL	**www.xyz.com**

Main Category	Subcategory	Examples	Normalization
Date	DATE	D.O.B: **12/08/2931**	**2931-08-12**
		Collected: **22/02/2043** at:… **Now** having palliative TAH ...	**2043-02-22**
			2043-03-21

	TIME	… A/Prof B Charles **at 10:30am on 28/4/73**.	**2073-04-28T10:30**
		Dr C Kobe **at 4:16am on the 11th of Oct 2713**	**2713-10-11T04:16**
	DURATION	… whilst **24weeks** pregnant in 2233	**P24W**
		Lung cancer resected **two weeks** ago.	**P2W**
	SET	… tested **twice** with positive controls;	**R2**

The table illustrates the main sensitive health information (SHI) categories and their corresponding subcategories, along with annotation examples, as defined in the annotation guidelines of the OpendDeID v2 corpus. Note that although annotation guideline [1] defines the "Profession" category, no annotations for this category were found during corpus construction. Consequently, the compiled corpus included six main SHI categories: NAME, LOCATION, AGE, DATE, CONTACT, and ID. Except for AGE, each category contained subcategories.

Supplementary Table 2. Statistics of the training, validation, and test sets of the OpenDeID v2 corpus.

Main Category	Sub Category	Training	Validation	Test	Total
Name	PATIENT	1,819	545	716	3080
	DOCTOR	6,541	2,191	3,327	12,059
Location	ROOM	1	0	0	1
	DEPARTMENT	1,025	366	419	1,810
	HOSPITAL	1,801	593	1,198	3,592
	ORGANIZATION	135	24	74	233
	STREET	899	321	344	1,564
	CITY	939	337	373	1,649
	STATE	875	310	332	1,517
	COUNTRY	3	2	0	5
	ZIP	913	325	353	1,591
	LOCATION-OTHER	6	4	6	16
Age		127	57	51	235
Date	DATE	4,753	1,622	2,459	8,834
	TIME	1,119	420	470	2,009
	DURATION	23	6	12	41
	SET	14	0	5	19
Contact	PHONE	9	2	1	12
	URL	3	4	3	10
IDs	MEDICALRECORD	1,824	563	747	3,134
	IDNUM	3,677	1,177	2,120	6,974

In the SREDH/AI CUP 2023 competition, the OpenDeID v2 corpus was divided into training, validation, and test sets consisting of 1,734, 560, and 950 reports, respectively. This table summarises the corpus in terms of the number of SHI annotations in the three datasets.

Supplementary Table 3. Performance comparison of full and LoRA fine-tuned models on the test set in terms of precision (P) and recall (R) scores.

	Subtask 1				Subtask 2			
	Micro		Macro		Micro		Macro	
	R	P	R	P	R	P	R	P
70M-Full	**0.566**	**0.901**	0.473	**0.739**	*0.668*	***0.790***	***0.398***	***0.788***
70M-LoRA	0.135	0.182	0.156	0.254	***0.759***	*0.187*	0.079	0.253
160M-Full	0.456	0.867	0.435	0.751	*0.649*	***0.778***	*0.441*	***0.823***
160M-LoRA	***0.833***	0.881	***0.644***	0.648	***0.743***	*0.637*	0.227	0.201
410M-Full	0.494	0.883	0.439	0.782	*0.634*	*0.762*	*0.345*	0.686
410M-LoRA	***0.899***	**0.904**	***0.773***	0.704	***0.839***	***0.767***	***0.911***	0.561
1B-Full	0.513	0.897	0.475	0.789	*0.636*	***0.795***	*0.436*	***0.757***
1B-LoRA	***0.880***	0.903	***0.761***	0.706	***0.770***	*0.668*	***0.708***	0.468
1.4B-Full	0.474	0.881	0.432	0.723	*0.585*	***0.786***	*0.325*	***0.750***
1.4B-LoRA	***0.893***	0.907	***0.782***	0.701	***0.742***	*0.655*	***0.740***	0.424
2.8B-Full	*0.774*	0.877	0.518	0.678	*0.696*	***0.783***	*0.331*	0.374
2.8B-LoRA	***0.906***	***0.925***	***0.850***	0.744	***0.804***	*0.722*	***0.866***	0.507
6.9B-Full	0.574	0.806	0.376	0.605	*0.599*	***0.796***	0.205	0.331
6.9B-LoRA	***0.874***	**0.902**	***0.849***	0.723	***0.815***	*0.728*	***0.867***	0.520
12B-Full	*0.754*	0.498	0.502	0.290	0.505	*0.803*	0.152	0.216
12B-LoRA	***0.912***	**0.824**	***0.751***	**0.649**	***0.751***	***0.827***	***0.600***	***0.897***
ICL-2.8B	0.611	0.906	0.498	0.824	0.338	0.580	0.247	0.678
ICL-6.9B	0.593	0.900	0.497	**0.827**	**0.398**	**0.630**	**0.265**	**0.691**
ICL-12B	**0.667**	**0.908**	**0.514**	0.784	0.384	0.583	0.211	0.681
Median	0.784	0.838	0.649	0.784	0.626	0.740	0.273	0.528

This table summarises the detailed P and R scores of the fine-tuned Pythia models. The Scores in bold indicate the highest performance achieved by the models of the same scale. Scores in italics indicate instances in which fine-tuned models outperformed the performance of the corresponding in-context learning (ICL) models. The underlined scores depict instances where fine-tuned models surpassed the median value of the competition.

Supplementary Table 4. Detailed performance of Team 1.

Coding Type	P	R	F	P	R	F
DATE	0.9888	0.9557	0.9719	0.8740	0.8211	0.8467
TIME	0.9896	0.8614	0.9211	0.8542	0.7979	0.8251
DURATION	0.7880	1.0000	0.8814	1.0000	0.8333	0.9091
SET	0.9944	0.9902	0.9923	1.0000	0.8000	0.8889
DOCTOR	0.9931	0.9394	0.9655			
IDNUM	0.8810	0.8773	0.8791			
MEDICALRECORD	0.9955	0.9340	0.9638			
PATIENT	0.9943	0.9276	0.9598			
CITY	0.9855	0.9884	0.9869			
DEPARTMENT	1.0000	0.9745	0.9871			
HOSPITAL	1.0000	0.9608	0.9800			
STATE	0.9280	0.8616	0.8936			
STREET	0.8333	0.8333	0.8333			
ZIP	0.9630	0.7027	0.8125			
AGE	0.8000	0.8000	0.8000			
ORGANIZATION	1.0000	0.3333	0.5000			
LOCATION-OTHER	0.9888	0.9557	0.9719			
PHONE	0.9896	0.8614	0.9211			

Supplementary Table 5. Detailed performance of Team 2.

Coding Type	P	R	F	P	R	F
DATE	0.8829	0.9752	0.9268	0.8249	0.8044	0.8145
TIME	0.8870	0.3340	0.4853	0.7325	0.2447	0.3668
DURATION	0.9231	1.0000	0.9600	1.0000	1.0000	1.0000
SET	1.0000	0.8000	0.8889	1.0000	0.8000	0.8889
DOCTOR	0.9785	0.9567	0.9674			
IDNUM	0.9836	0.9887	0.9861			
MEDICALRECORD	0.7880	1.0000	0.8814			
PATIENT	0.9916	0.9930	0.9923			
CITY	0.9919	0.9839	0.9879			
DEPARTMENT	0.9369	0.9212	0.9290			
HOSPITAL	0.9094	0.9641	0.9360			
STATE	0.9970	0.9970	0.9970			
STREET	0.9913	0.9971	0.9942			
ZIP	0.9971	0.9971	0.9971			
AGE	0.9608	0.9608	0.9608			
ORGANIZATION	0.9859	0.9459	0.9655			
LOCATION-OTHER	1.0000	0.5000	0.6667			
PHONE	1.0000	1.0000	1.0000			

Supplementary Table 6. Detailed performance of Team 3.

Coding Type	P	R	F	P	R	F
DATE	0.985	0.9101	0.9461	0.8275	0.7532	0.7886
TIME	0.9461	0.934	0.94	0.8519	0.7957	0.8229
DURATION	0.8182	0.75	0.7826	1.0	0.75	0.8571
SET	1.0	0.8	0.8889	1.0	0.8	0.8889
DOCTOR	0.9517	0.9591	0.9554			
IDNUM	0.8837	0.9783	0.9286			
MEDICALRECORD	0.7863	1.0	0.8804			
PATIENT	0.8825	0.9651	0.9219			
CITY	0.9748	0.933	0.9534			
DEPARTMENT	0.9196	0.8735	0.896			
HOSPITAL	0.878	0.8531	0.8654			
STATE	0.991	0.991	0.991			
STREET	0.9799	0.9942	0.987			
ZIP	0.9915	0.9858	0.9886			
AGE	0.92	0.902	0.9109			
ORGANIZATION	0.7536	0.7027	0.7273			
LOCATION-OTHER	1.0	0.1667	0.2857			
PHONE	0.0	0.0	0.0			

Supplementary Table 7. Detailed performance of Team 4.

Coding Type	P	R	F	P	R	F
DATE	0.8657	0.9699	0.9148	0.8143	0.7898	0.8018
TIME	0.5753	0.2277	0.3262	0.7453	0.1681	0.2743
DURATION	0.8000	1.0000	0.8889	1.0000	1.0000	1.0000
SET	0.6667	0.8000	0.7273	1.0000	0.8000	0.8889
DOCTOR	0.9783	0.9098	0.9428			
IDNUM	0.9827	0.7241	0.8338			
MEDICALRECORD	0.7886	0.9987	0.8813			
PATIENT	0.9916	0.9902	0.9909			
CITY	0.9780	0.9517	0.9647			
DEPARTMENT	0.9052	0.8663	0.8854			
HOSPITAL	0.8793	0.8940	0.8866			
STATE	0.9754	0.9548	0.9650			
STREET	1.0000	0.9971	0.9985			
ZIP	0.9971	0.9773	0.9871			
AGE	0.9722	0.6863	0.8046			
ORGANIZATION	0.9444	0.9189	0.9315			
LOCATION-OTHER	1.0000	0.5000	0.6667			
PHONE	1.0000	1.0000	1.0000			

Supplementary Table 8. Detailed performance for Team 5

Coding Type	P	R	F	P	R	**F**
DATE	0.8898	0.9723	0.9293	0.8210	0.7983	0.8095
TIME	0.8835	0.4681	0.6120	0.8318	0.3894	0.5304
DURATION	0.7143	0.8333	0.7692	1.0000	0.8333	0.9091
SET	1.0000	0.8000	0.8889	1.0000	0.8000	0.8889
DOCTOR	0.9256	0.9423	0.9339			
IDNUM	0.8907	0.8958	0.8932			
MEDICALRECORD	0.7697	0.9799	0.8622			
PATIENT	0.9916	0.9902	0.9909			
CITY	0.9559	0.9303	0.9429			
DEPARTMENT	0.8525	0.8138	0.8327			
HOSPITAL	0.8340	0.8139	0.8238			
STATE	0.9760	0.9819	0.9790			
STREET	1.0000	0.9971	0.9985			
ZIP	0.9971	0.9773	0.9871			
AGE	0.7903	0.9608	0.8673			
ORGANIZATION	0.6162	0.8243	0.7052			
LOCATION-OTHER	1.0000	0.3333	0.5000			
PHONE	0.0000	0.0000	0.0000			

Supplementary Table 9. Detailed performance for Team 6

Coding Type	P	R	F	P	R	F
DATE	0.9712	0.9199	0.9449	0.8209	0.7552	0.7867
TIME	0.9521	0.8872	0.9185	0.7482	0.6638	0.7035
DURATION	0.7500	0.5000	0.6000	1.0000	0.5000	0.6667
SET	1.0000	0.6000	0.7500	1.0000	0.6000	0.7500
DOCTOR	0.9805	0.9375	0.9585			
IDNUM	0.8025	0.8759	0.8376			
MEDICALRECORD	0.7565	0.8983	0.8213			
PATIENT	0.9684	0.9413	0.9547			
CITY	0.9694	0.9357	0.9523			
DEPARTMENT	0.8690	0.8234	0.8456			
HOSPITAL	0.8913	0.8623	0.8765			
STATE	0.9909	0.9849	0.9879			
STREET	0.9603	0.8430	0.8978			
ZIP	0.9915	0.9887	0.9901			
AGE	0.9167	0.8627	0.8889			
ORGANIZATION	0.7703	0.7703	0.7703			
LOCATION-OTHER	1.0000	0.5000	0.6667			
PHONE	1.0000	1.0000	1.0000			

Supplementary Table 10. Detailed performance for Team 7

Coding Type	P	R	F	P	R	F
DATE	0.9661	0.9626	0.9644	0.8217	0.7910	0.8061
TIME	0.4569	0.6085	0.5219	0.8339	0.4809	0.6100
DURATION	0.9167	0.9167	0.9167	1.0000	0.9167	0.9565
SET	1.0000	0.8000	0.8889	1.0000	0.8000	0.8889
DOCTOR	0.9496	0.9459	0.9477			
IDNUM	0.9651	0.9906	0.9777			
MEDICALRECORD	0.7880	1.0000	0.8814			
PATIENT	0.9764	0.9818	0.9791			
CITY	0.9595	0.9517	0.9556			
DEPARTMENT	0.8658	0.7542	0.8061			
HOSPITAL	0.7277	0.8698	0.7924			
STATE	0.9642	0.9729	0.9685			
STREET	0.9884	0.9913	0.9898			
ZIP	0.9887	0.9943	0.9915			
AGE	0.9111	0.8039	0.8542			
ORGANIZATION	0.5200	0.5270	0.5235			
LOCATION-OTHER	0.0000	0.0000	0.0000			

PHONE	0.0000	0.0000	0.0000			

Supplementary Table 11. Detailed performance for Team 8

Coding Type	P	R	F	P	R	F
DATE	0.9850	0.9101	0.9461	0.8275	0.7532	0.7886
TIME	0.9461	0.9340	0.9400	0.8519	0.7957	0.8229
DURATION	0.8182	0.7500	0.7826	1.0000	0.7500	0.8571
SET	1.0000	0.8000	0.8889	1.0000	0.8000	0.8889
DOCTOR	0.9517	0.9591	0.9554			
IDNUM	0.8837	0.9783	0.9286			
MEDICALRECORD	0.7863	1.0000	0.8804			
PATIENT	0.8825	0.9651	0.9219			
CITY	0.9748	0.9330	0.9534			
DEPARTMENT	0.9196	0.8735	0.8960			
HOSPITAL	0.8780	0.8531	0.8654			
STATE	0.9910	0.9910	0.9910			
STREET	0.9799	0.9942	0.9870			
ZIP	0.9915	0.9858	0.9886			
AGE	0.9200	0.9020	0.9109			
ORGANIZATION	0.7536	0.7027	0.7273			
LOCATION-OTHER	1.0000	0.1667	0.2857			
PHONE	0.0000	0.0000	0.0000			

Supplementary Table 12. Detailed performance for Team 9

Coding Type	P	R	F	P	R	**F**
DATE	0.8341	0.9776	0.9002	0.8265	0.8081	0.8172
TIME	0.3500	0.0298	0.0549	0.4286	0.0128	0.0248
DURATION	0.8000	1.0000	0.8889	1.0000	1.0000	1.0000
SET	1.0000	0.8000	0.8889	1.0000	0.8000	0.8889
DOCTOR	0.9659	0.9615	0.9637			
IDNUM	0.9763	0.9920	0.9841			
MEDICALRECORD	0.7886	0.9987	0.8813			
PATIENT	0.9916	0.9930	0.9923			
CITY	0.9651	0.9625	0.9638			
DEPARTMENT	0.8929	0.8950	0.8939			
HOSPITAL	0.8778	0.8932	0.8854			
STATE	0.9639	0.9639	0.9639			
STREET	1.0000	0.9971	0.9985			
ZIP	1.0000	0.9972	0.9986			
AGE	0.9245	0.9608	0.9423			
ORGANIZATION	0.5443	0.5811	0.5621			

LOCATION-OTHER	0.7500	0.5000	0.6000			
PHONE	1.0000	1.0000	1.0000			

Supplementary Table 13. Detailed performance for Team 10

Coding Type	P	R	F	P	R	F
DATE	0.8900	0.9776	0.9318	0.8325	0.8003	0.8161
TIME	0.8341	0.8234	0.8287	0.8307	0.6787	0.7471
DURATION	0.6667	0.3333	0.4444	1.0000	0.3333	0.5000
SET	1.0000	0.8000	0.8889	1.0000	0.8000	0.8889
DOCTOR	0.9547	0.9050	0.9292			
IDNUM	0.9747	0.9821	0.9784			
MEDICALRECORD	0.7856	0.9960	0.8784			
PATIENT	0.9834	0.9902	0.9868			
CITY	0.9625	0.9625	0.9625			
DEPARTMENT	0.8504	0.9093	0.8789			
HOSPITAL	0.8840	0.9032	0.8935			
STATE	0.9819	0.9789	0.9804			
STREET	0.9684	0.9797	0.9740			
ZIP	1.0000	0.9320	0.9648			
AGE	0.9111	0.8039	0.8542			
ORGANIZATION	0.5979	0.7838	0.6784			
LOCATION-OTHER	1.0000	0.1667	0.2857			
PHONE	0.0000	0.0000	0.0000			

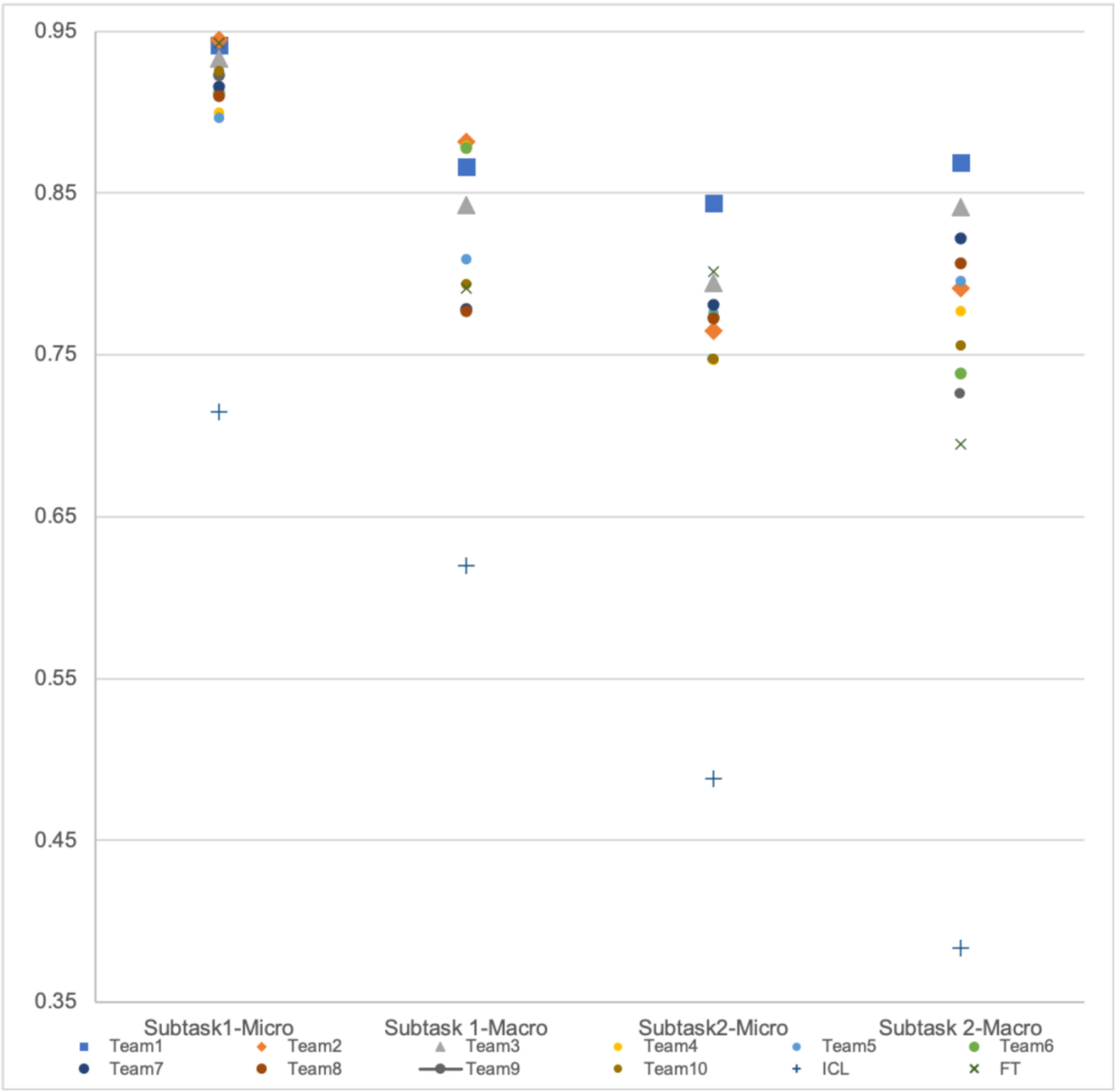
Supplementary Table 14. Expanded tokens for fine-tuning the pre-trained Pythia models

Special Token	Symbol
bos_token	<|endoftext|>
eos_token	<|END|>
pad_token	<|pad|>
sep_token	\n\n####\n\n

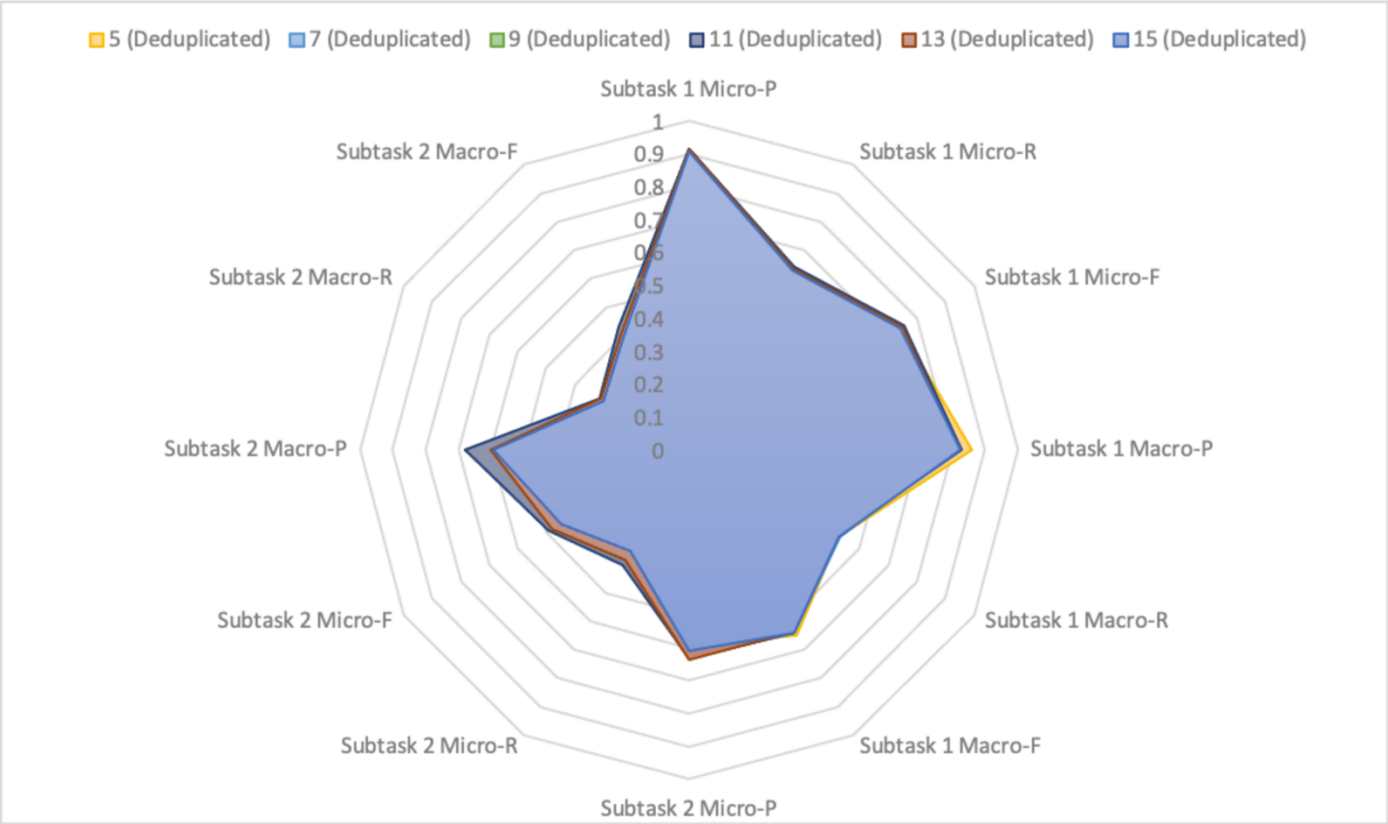
In our implementation, we extended Pythia's built-in tokenizer vocabulary with special symbols defined in the table. The expanded vocabulary contains 50,280 tokens. Because the lengths of the training instances may differ in a batch, pad_token was used to ensure that the lengths of the training instances in a batch were the same. The padded tokens are masked during the training phase to avoid the self-attention mechanism of the model paying attention on them.

Supplementary Figures

Supplementary Figure 1. Performance comparison for subtasks 1 and 2 based on micro- and macro-F-scores on the test sets. The figure presents the performance of the top 10 teams, comparing both micro- and macro-averaged F1 scores for the two subtasks alongside team names. In subtask 1, the top-ranked system developed by Huang et al. [2] utilised the LLM-based method based on Qwen developed by Bai et al. [3], thereby achieving micro- and macro-averaged F1 scores of 0.945 and 0.882, respectively. In subtask 2, the highest-ranking system relying on pattern-based approaches was developed by Zhao et al. [4] achieving micro- and macro-averaged F1 scores of 0.844 and 0.869, respectively. The overall performance showed average micro-/macro-F scores for subtasks 1 and 2 of 0.666/0.496 and 0.6/0.394, with median micro-/macro-F scores of 0.810/0.710 and 0.679/0.360, respectively.



Supplementary Figure 2. Impact of the number of in-context examples on the performance of the Pythia-2.8B model for Subtasks 1 and 2, based on the deduplicated training set. The figure illustrates the effect of the number of in-context examples on the performance of both subtasks when using a deduplicated training set. We followed the same experimental setup as in Figure 4, with the only difference being that all duplicated training instances were removed from the dataset before sampling examples for ICL. From the results, we observe that the optimal number of in-context examples for Subtask 1 remains at five. However, for Subtask 2, both precision and recall continued to improve as the number of examples increased, with the best performance achieved when selecting up to 11 examples. Compared to the results obtained using the original non-deduplicated training set, the deduplicated setup leads to improved performance, highlighting the benefits of increased contextual diversity and reduced redundancy in the ICL examples.



Supplementary Figure 3. Sample de-identified text report with SHIs highlighted in bold. The figure illustrates a segment of a de-identified report text in which the SHI is highlighted in bold. The text in the figure only contains synthetic SHIs; they have been de-identified, which means that the original true SHI mentions have been replaced by surrogate texts. Here, true SHIs are those present in the original text, whereas surrogate SHIs are substituted from one or more sources to replace the true SHIs. Furthermore, for date-related SHI types, such as date of birth/collection and specific weekdays like "Friday", they are derived from the original real dates by adding a random number of days according to pre-defined constraints [5] to maintain the time intervals between events accurately.

D.O.B: **09/08/2957**
Sex: F
Collected: **14/02/3014 at 11:42**
Location: **3 ARRIETTA CLOSE-POW**
DR **AADLAND ABRAHAM**
…
Result required for multidisciplinary meeting on **Friday**.

Supplementary Figure 4. Sample annotation file. The file shows an example annotation file for Supplementary Figure 3. Each line in the annotation file (Supplementary Figure 3) corresponds to a SHI instance. The first field indicates the file ID. The second field specifies the SHI type, followed by the start and end indices of the SHI mentioned in the text, which are recorded in the fifth field. The last field contains the normalised value, applicable if the instance belongs to the 'Date' main category.

9 DATE 8 18 09/08/2957 **2957-08-09**

9 TIME 38 57 14/02/3014 at 11:42 **3014-02-14T11:42**

9 DATE 208 214 Friday **3014-02-18**

9 DEPARTMENT 69 85 3 ARRIETTA CLOSE

9 HOSPITAL 86 89 POW

9 DOCTOR 93 108 AADLAND ABRAHAM

Supplementary Figure 5. The fine-tuned data template format and converted training example for the first training instance are shown in Supplementary Figure 3. This figure illustrates the tokenised input/output structure used in tagging tasks. The input sequence includes tokens {bos_token} and {sep_token}, which delineate the raw input text (CONTENT) from its corresponding label section (LABELS), followed by a {eos_token} to signal the end of the sequence.

{bos_token}{CONTENT}{sep_token}{LABELS}{eos_token}
<|endoftext|> D.O.B: 09/08/2957 #### DATE: 09/08/2957=>2957-08-09<|END|>

Supplementary Note 1. Evaluation metrics

We used precision, recall, and micro/macro-F_1 measures to evaluate and compare participants' predictions on the validation and final test sets against gold annotations for both subtasks. For Subtask 1, the three fields of an SHI instance, the start index, end index, and SHI type, were considered to assess the number of true positives (TP), false positives (FP), and false negatives (FN). For Subtask 2, the normalised temporal expression value was considered in addition to the three fields mentioned above. Based on the characteristics of the predicted SHIs, a TP was defined as accurately extracting all three or four fields (depending on the task type), as provided by human annotators for gold SHI annotation. An FP occurs when any field of a predicted SHI does not match that of manual annotations. Conversely, an FN case refers to a manually annotated SHIs without exact matching model prediction.

The formulas for the precision (P), recall (R), and F_1 score for a specific SHI subcategory c are as follows:

$$\mathrm{P}_{(c)} = \frac{\mathrm{TP}_{(c)}}{\left(\mathrm{TP}_{(c)}+\mathrm{FP}_{(c)}\right)} \tag{1}$$

$$\mathrm{R}_{(c)} = \frac{\mathrm{TP}_{(c)}}{\left(\mathrm{TP}_{(c)}+\mathrm{FN}_{(c)}\right)} \tag{2}$$

$$\text{Micro-}\mathrm{F1}_{(c)} = 2 \times \frac{\left(\mathrm{P}_{(c)}\times\mathrm{R}_{(c)}\right)}{\left(\mathrm{P}_{(c)}+\mathrm{R}_{(c)}\right)} \tag{3}$$

$$\text{Macro-avg } \mathrm{F1}_{(c)} = \frac{\sum_{c\in C} \text{Micro-}\mathrm{F1}_{(c)}}{|C|} \tag{4}$$

where C is the set of all the SHI subcategories. In the SREDH/AICUP 2023 competition, participants' predictions for both subtasks were assessed and ranked using the macro-averaged F_1 score, as shown in Supplementary Formula 4. Final rankings were determined by averaging the individual ranks from both tasks for each participant, and sorting the teams based on the averaged values.

Supplementary Note 2. Model development details

The training data introduced in Supplementary Figure 4 were tokenised using the built-in tokenizer of a pretrained Pythia model, a suite of autoregressive language models (LMs) [6]. The HuggingFace Transformer implementation GPTNeoXTokenizerFast was used to process the dataset. In our implementation, we extended the tokenizer vocabulary (original size of 50,277 tokens) by including the special symbols defined in Supplementary Table 14.

In addition to full-parameter fine-tuning of all Pythia models, we applied low-rank adaptation (LoRA) [7] to develop another set of fine-tuned models. In this approach, the parameters of the pre-trained Pythia models were frozen, and only the trainable adapter parameters applied to the query, key, and value matrices within the transformer attention mechanism were optimised. The LoRA rank r was set to eight for all LoRA fine-tuned models. During the fine-tuning process, the number of epochs was set to 200 by using an early stop strategy on the validation set to prevent overfitting. If there was no progress in the validation set, the number of epochs to wait before the early stop was set to three. The batch size was adjusted to maximise the GPU utilisation, and the learning rate was set to 10^{-4}.

Supplementary Note 3. Strategies employed by the top four teams

A brief summary of the strategies employed by the top four teams is presented below:

Team 1 conducted a comparative study to evaluate the efficacy of two approaches: fine-tuning the ChatGPT (based on the GPT-3.5-turbo-1106 model) and a rule-based approach. For the LLM-based method, the team preprocessed the training set by segmenting the sentences and associating them with the corresponding annotations to generate paragraph–annotation pairs for instruction tuning, following the specifications of the ChatGPT API. The length of each segmented paragraph was limited to 4,096 tokens to meet the API requirements. After initially fine-tuning the model for three epochs, the team adopted a sampling approach to select sentences from the entire training set that contained SHI categories with fewer than 150 instances. The initial fine-tuned model then underwent a second phase of training on these sampled data with the goal of improving the performance of these underrepresented categories.

In contrast, the rule-based approach involves compiling a set of rules and dictionaries collected from online resources and publicly available datasets to recognize and normalize SHIs. The team developed rules via regular expressions and employed an analysis tool that the team developed to provide a comparative function to assess the disparities between the gold standard training set and the results predicted from the compiled rules. The tool categorizes disparities as false positives or false negatives, which are then manually reviewed to refine or modify the patterns until the desired performance is achieved. For SHIs in the ORGANIZATION category, the team found that regular expressions were ineffective; therefore, the team used a dictionary lookup method with precompiled lexicons. For Subtask 2, the team developed normalization rules tailored to each date-related SHI subcategory.

Team 2 utilized the Qwen-14B and 7B models for subtasks 1 and 2, respectively, based on the findings of Wang, et al. [82] and Zhu, et al. [83], indicating that the Qwen model performs well in handling clinical notes. To enhance training stability, the team implemented a fully sharding data parallel technique [84]. To mitigate potential overfitting, the team employed the NEFTune method [85], which introduces uniformly distributed noise to the embedding layer of the model. This noise prevents the model from memorizing specific details within the training set, thereby allowing it to generalize more effectively to new data. Additionally, this technique reduces the model's sensitivity to specific inputs, preventing it from developing overly complex representations too early in the training process. Another strategy employed by the team was to randomly extract sentences within a predefined context window and use them as the basis for validating the model's performance. The length of the context window was set to 1 for subtask 1 but 3 for subtask 2, as the normalization of temporal information often requires more contextual data than the other subtasks.

During the inference phase, the team used a greedy decoding strategy to ensure the stability and reliability of the model output. In addition, the team incorporated a rule-based postprocessing method since the team observed that the fine-tuned model occasionally makes erroneous predictions for some labels that should be straightforward to classify.

Team 6 employed a Longformer model in combination with conditional random fields (CRFs) to tackle the challenge. The Longformer was chosen for its ability to process up to 4096 tokens, enabling it to capture extensive context in EHRs. Several preprocessing steps were applied, including tokenization and annotation verification. The team employed a sliding window technique to manage texts that exceeded the token limit, thereby ensuring continuity across segments.

For model training, the team incorporated a fully connected dense layer with 768 units after the transformer output of the Longformer, followed by a CRF layer to improve the sequence labeling. The dense layer captures rich contextual information, whereas the CRF layer enhances the model's ability to recognize patterns in which certain SHI entities tend to appear consecutively. This approach emphasizes the importance of a large context [86]. Furthermore, the team implemented a 5-fold ensemble learning approach in which predictions from five distinct models were aggregated using a voting system. This method resulted in a final macro-F1 score of 0.878 on the test set. Following the initial model implementation, the team integrated a rule-based normalization strategy. Using regular expressions and word-to-number conversion, the team normalized the textual numbers and date formats.

The method proposed by team 8 involves preprocessing, model implementation, and rule-based data extraction. The preprocessing steps include text segmentation, prompt generation to extract HIPAA-related information, and temporal expression normalization. The processed data were formatted to meet the requirements of the Hugging Face dataset library, which enabled seamless integration with advanced learning algorithms. The outcomes were serialized in the JSONL format, and postprocessing steps were employed to refine the extracted information, ensuring compatibility with downstream tasks.

For text generation, the team utilized the T5-Efficient-BASE-DL2 model [87], which is an optimized version of the T5 architecture that offers heightened computational efficiency and reduced memory usage. The model was trained on the preprocessed data via serialized JSONL input-output pairs, with hyperparameters such as a learning rate of 2e--5, a batch size of 4, and 10 training epochs. During training, the team divided the released training dataset into training, validation, and test sets. The team subsequently merged the training and validation sets for final tuning, utilized a 90-10 split for training and validation, and employed Hugging Face's Seq2SeqTrainer to manage the training process efficiently.

In addition to model-based approaches, rule-based methods have been developed to extract identifiable information, such as phone numbers, locations, and city names. These rules included customized regular expressions for local formatting and geographic identifiers, along with a temporal normalization strategy that followed ISO 8601 standards. Temporal expressions such as "two weeks" or "three months were systematically converted into structured representations (e.g., "P2W" for two weeks) to ensure consistency and facilitate further analysis. This combination of model-based text generation and rule-based extraction results in a robust framework for deidentification that can effectively handle both structured and unstructured data.

References

1 Dai, H.-J. & Jonnagaddala, J. *HSA Study PHI Corpus - Annotation guidelines* <https://nkustislab.github.io/docs/project/aicup/guideline.pdf> (2023).

2 Huang, C.-L., Rianto, B., Sun, J.-T., Fu, Z.-X. & Lee, C.-H. in *Proceedings of the 2024 International Workshop on Deidentification of Electronic Medical Record Notes* (Springer Nature, Kaohsiung, Taiwan, 2024).

3 Bai, J. *et al.* Qwen technical report. *arXiv preprint arXiv:2309.16609* (2023).

4 Zhao, Z.-R., Chou, P.-C., Mir, T. H. & Dai, H.-J. in *Proceedings of the 2024 International Workshop on Deidentification of Electronic Medical Record Notes* 27-38 (Springer Nature, Kaohsiung, Taiwan, 2024).

5 Saeed, M., Lieu, C., Raber, G. & Mark, R. G. in *Computers in cardiology.* 641-644 (IEEE).

6 Biderman, S. *et al.* in *International Conference on Machine Learning.* 2397-2430 (PMLR).

7 Hu, E. J. *et al.* in *International Conference on Learning Representations.*