S1. Annotation Guidelines

The task involved removing Personally Identifiable Information (PII) from CHAMPS infant mortality reports. The manually reviewed and de-identified subset contained PII limited to personal names, geograPIIc locations, and dates. To ensure consistency, multi-word PII spans were replaced with single redaction tokens (e.g., January 1st was replaced with `<date>`).

S1.1. Location

1. Names of medical facilities were also removed, but the facility type—such as hospital, health centre, or dispensary—was retained to preserve contextually relevant information.
2. In cases where the proper name of a facility was interrupted by the facility type (e.g., Hospital-name Hospital at Town), the redaction preserved the surrounding structure using multiple tokens (e.g., <location> hospital at <location>).
3. Standard hospital unit names such as ICU were not de-identified.
4. In contrast, names of government and private hospitals were redacted due to the potential for re-identification in settings with a single such institution.

Table S1: Location de-identification examples.

Original Example	**Deidentified Example**
xyz hospital xyz	<location> hospital <location>
xyz	<location>
Bob hosp.	<location> hosp.
the xyz Hospital	the <location> Hospital
xyz Health Centre	<location> Health Centre
xyz Referral	<location>
nearby private clinic	nearby <location> clinic
The private doctor	The <location> doctor
A private clinic	A <location> clinic
Local clinic	<location> clinic
Favor xyz	<location>
Healthcare/medicinal facilities of different types	<location> medical center
xyz clinic	<location> clinic
xyz health clinic	<location> health clinic
xyz dispensary	<location> dispensary
xyz pharmacy	<location> pharmacy
Dr. xyz clinic	<location> clinic

Table S2: Common locations and region-specific names do not need to be removed.

Content	**Example**
Common hospital unit names	ANC, NICU, SCBU, KMC, RESUS
Region-specific names for vehicles	okada, boda boda, tuk tuk, matatu
Regions-specific clinic types	CHC, CHP, PHU
Mentions of CDC/CHAMPS/country name	CDC, CHAMPS, Nigeria
Words for nursery/daycare	Creche
Region specific words	Manyasi (an herb mixture)

Ward names were removed in case they were specific to the hospital (e.g., “ward 66”).

S1.2. Date

1. The day and month were removed, while the year was retained to preserve temporal context without compromising privacy.
2. Standalone days and/or months were also removed.

Table S3: Date de-identification examples.

Monday	Monday
January 2019	<date> 2019
saturday, june 2024	saturday, <date> 2024
26th of may 2023	<date> 2023
14/06/2023	<date>/2023
26th of December	<date>
the 26th of December	the <date>
the 3rd and 4th of February 2024	the <date> and <date> 2024
On the 8$^{\text{th}}$ of the month	On the <date> of the month

Table S4: Content which should NOT be removed.

Content	**Example**
Age/Gestational age (months or weeks and days)	35 yrs old, baby born in 8 weeks
Time (hours/minutes)	1pm, 11:00

S1.3. Names

All personal names were removed, including those of infants, parents, medical providers, and other identified individuals.

Table S5: Name de-identification examples.

xyz's death	<name>'s death

S2. Iterative Prompt Engineering for PII Extraction

To optimize the performance of the LLM in identifying and handling PII within unstructured clinical narratives, an iterative prompt engineering strategy was employed. The prompt architecture progressed through three distinct developmental phases, transitioning from zero-shot inline redaction to dynamic, retrieval-augmented few-shot extraction. The objective of this iterative process was to improve extraction accuracy, reduce hallucinations, and enforce strict output formatting for downstream evaluation.

S2.1. Baseline Zero-Shot Redaction

The initial baseline approach (Prompt 1) utilized a zero-shot framework instructing the LLM to act as a medical assistant. The primary task was formulated as a direct text transformation: the model was asked to read the clinical note and generate a fully de-identified version by replacing specific PII with generalized tags. The prompt explicitly defined the scope of the task by categorizing the target PII into three specific types: Name (patients and employers), Location (including government hospitals), and Date (specifically day and month, excluding weekdays and years). To guide the model, brief, inline text examples were provided, demonstrating how specific entities should be replaced by generic tokens (e.g., replacing "xyz hospital xyz" with <location> hospital <location>). While conceptually straightforward, instructing the model to rewrite the entire text sequence increases the risk of unintended formatting alterations and complicates automated performance evaluation.

```
You are a medical assistant. Read the clinical note carefully. Remove all Protected
    Health Information (PHI) from the clinical note and output a deidentified version
    of the clinical note. Preserve all non-PHI text as they appear in the original text
    . Do not standardize, reformat, or alter the text in any way.

Types of PHI to identify:
1. Name: Specific names of patients and employers.
2. Location: Specific locations including a government hospital.
3. Date: Dates containing a specific day and month. Does not count weekdays and year.

Examples:
Input: ###START nightingale hospital kondele END###
Output: ###START <location> hospital <location> END###

Input: ###START Bob hosp. END###
Output: ###START <location> hosp. END###

Input: ###START January 2019 END###
Output: ###START <date> 2019 END###

Input: ###START 14/06/2023 END###
Output: ###START <date>/2023 END###

Input:###START {{text}} END###
```

Prompt 1: Baseline Zero-Shot Redaction Prompt

S2.2. Structured Zero-Shot Extraction

To address the limitations of the baseline redaction approach, the second iteration (Prompt 2) shifted the core objective from text rewriting to structured entity extraction. The instructions were refined to demand that the model identify PII and output it strictly as key-value pairs (e.g., <PII type>=PII terms). To enhance the model's comprehension of complex de-identification rules—particularly regarding the partial extraction of dates and specific hospital names—the formatting of the in-context examples was significantly upgraded. Static examples were presented utilizing structured Markdown tables, delineating the "Original Example" alongside the "Deidentified Example." This structural constraint was designed to force the LLM to focus solely on the precise spans of text requiring extraction, thereby minimizing the generation of extraneous text and ensuring a standardized output format that is easily parsable for algorithmic scoring.

```
You are a medical assistant. Your task is to identify all Protected Health Information
     (PHI) from the clinical note as below.

Types of PHI to identify:
1. Name: Specific names of patients and employers.
2. Location: Specific locations including a government hospital.
3. Date: Dates containing a specific day and month. Does not count weekdays and year.

Examples:
### Hospital Deidentification
| Original Example                  | Deidentified Example              |
|-----------------------------------|-----------------------------------|
| xyz hospital xyz                  | <location>=xyz, xyz               |
| xyz                               | <location>=xyz                    |
| Xyz hosp.                         | <location>=Xyz                    |
| the Xyz Hospital                  | <location>=Xyz                    |

---
```

```
### Date Deidentification Rules
| Original Example                  | Deidentified Example              |
|-----------------------------------|-----------------------------------|
| January 2019                      | <date>=January                    |
| saturday, june 2024               | <date>=june                       |
| 26th of may 2023                  | <date>=26th of may                |
| 14/06/2023                        | <date>=14/06                      |
| the 26th of December              | <date>=26th of December           |

Preserve all PHI terms as they appear in the original text. Do not standardize,
    reformat, or alter the terms in any way. Your answer should be formatted as "<PHI
    type>=PHI terms". For example:

"""
<NAME>=John Doe
<LOCATION>=123 Main St
"""

Clinical note:
{{text}}

Please provide your answer below:
```

Prompt 2: Structured Zero-Shot Extraction Prompt

S2.3. Dynamic Few-Shot Prompting via Retrieval

The final iteration transitioned from relying on static, zero-shot instructions to a dynamic, few-shot paradigm. Recognizing the high linguistic variability inherent in clinical narratives, a Retrieval-Augmented Generation (RAG) inspired pipeline was implemented to provide the LLM with contextually relevant exemplars (Prompt 3). This approach involved a three-step pipeline:

- Index Construction: A search engine was built using a training corpus of previously annotated clinical notes.
- Semantic Retrieval: For each novel clinical note presented during inference, the system queried the index to retrieve the top N most semantically similar notes from the training set.
- Prompt Augmentation: The retrieved clinical notes, along with their corresponding gold-standard extracted PII pairs, were dynamically appended to the prompt instructions.

However, contrary to initial hypotheses, empirical evaluation demonstrated that the inclusion of retrieved few-shot exemplars did not yield a measurable improvement in overall system performance. The added complexity of the retrieval pipeline, potentially introducing distracting or misaligned context, ultimately failed to enhance the LLM's extraction capabilities beyond the baseline established in the previous iteration. Consequently, the dynamic few-shot approach was discarded in favor of the structured zero-shot extraction methodology, which provided the optimal balance of accuracy, formatting consistency, and computational efficiency for the final system architecture.

```
You are a medical assistant. Your task is to identify all Protected Health Information
     (PHI) from the clinical note as below.
Types of PHI to identify:
1. Name: Specific names of patients and employers.
2. Location: Specific locations including a government hospital.
3. Date: Dates containing a specific day and month. Does not count weekdays and year.
Preserve all PHI terms as they appear in the original text. Do not standardize,
    reformat, or alter the terms in any way. Your answer should be formatted as "<PHI
    type>=PHI terms".

Examples:
```

```
Clinical note:
I missed my periods in December 2022. I bought a pregnancy test in January 2023 and it
     showed two lines even though one was faded. I started my ANC in February 2023, I
    was given folic acid and ferrous sulphate. I was admitted for a high BP and the
    baby was getting tired. I was booked for an emergency caesarian section. The baby
    was born with a lung problem. They said his lungs had air and a tube was inserted
    to drain the air. He also had a heart problem, they said he had a bleed in his
    brain beca use of prematurity. The baby was brain damaged and he died.
Answer:
<date>=december, january, february

Clinical note:
{{text}}

Please provide your answer below:
```

Prompt 3: Dynamic Few-Shot Prompting with Retrieved Exemplars

S3. Evaluation Metrics

We evaluated system performance against a manually-annotated gold standard, excluding the samples used for guideline preparation and system development. Evaluation was performed at the entity-token level; each instance of a true PII entity was independently assessed. An entity tag was considered incorrect if any token within the relevant span was misclassified. Based on this strict span requirement, we defined our error categories as follows:

- **False Negative (FN):** Occurs if any portion of a true PII entity remained unredacted (a privacy leak).
- **False Positive (FP):** Occurs if any non-PII text was incorrectly redacted (over-redaction).
- **True Positive (TP):** Occurs only when the relevant span is correctly identified and classified.

We measured performance using Precision, Recall, and F_1-score. **Precision** captures the extent of over-redaction (the reliability of positive predictions) and is calculated as:

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{S1}$$

Recall measures the system's success at preventing data leaks (the coverage of true PII entities) and is calculated as:

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{S2}$$

The $\mathbf{F_1}$**-score** provides the harmonic mean of the two, offering a balanced view of system performance:

$$\text{F}_1 = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{S3}$$

S4. Detailed Performance Metrics

Table S6: Performance of the de-identification system (named entity recognition) on the annotated data, grouped by dataset and entity type. The table displays Precision, Recall, and F_1-score with 95% confidence intervals for each country.

Entity	Country	Precision (95% CI)	Recall (95% CI)	F_1-score (95% CI)
Verbal Autopsies (VA)				
Date	Kenya	1.00 (1.00-1.00)	1.00 (1.00-1.00)	1.00 (1.00-1.00)
	Sierra Leone	0.98 (0.97-1.00)	1.00 (0.97-1.00)	0.99 (0.97-1.00)
	South Africa	1.00 (0.93-0.96)	0.89 (0.93-0.96)	0.94 (0.93-0.96)
	Overall	0.98 (0.97-0.99)	0.93 (0.91-0.95)	0.96 (0.95-0.97)
Location	Ethiopia	0.75 (0.70-0.80)	0.98 (0.96-1.00)	0.85 (0.81-0.88)
	Kenya	1.00 (1.00-1.00)	1.00 (1.00-1.00)	1.00 (1.00-1.00)
	Sierra Leone	0.90 (0.87-0.93)	1.00 (0.99-1.00)	0.94 (0.93-0.96)
	South Africa	0.97 (0.95-0.98)	0.97 (0.95-0.99)	0.97 (0.96-0.98)
	Overall	0.90 (0.88-0.91)	0.99 (0.98-0.99)	0.94 (0.93-0.95)
Name	Ethiopia	0.51 (0.43-0.59)	1.00 (1.00-1.00)	0.67 (0.60-0.74)
	Kenya	0.83 (0.60-1.00)	1.00 (1.00-1.00)	0.91 (0.75-1.00)
	Sierra Leone	0.65 (0.50-0.79)	1.00 (1.00-1.00)	0.79 (0.67-0.89)
	South Africa	0.86 (0.75-0.96)	0.93 (0.84-1.00)	0.89 (0.82-0.96)
	Overall	0.61 (0.55-0.67)	0.98 (0.96-1.00)	0.75 (0.70-0.80)
Maternal Clinical Abstractions (MCA)				
Date	Bangladesh	0.99 (0.99-1.00)	0.99 (0.99-1.00)	0.99 (0.99-1.00)
	Ethiopia	0.99 (0.95-1.00)	1.00 (1.00-1.00)	0.99 (0.98-1.00)
	Kenya	0.97 (0.96-0.98)	0.98 (0.97-0.99)	0.98 (0.97-0.99)
	Mozambique	1.00 (1.00-1.00)	0.83 (0.65-1.00)	0.91 (0.77-1.00)
	Nigeria	1.00 (1.00-1.00)	0.97 (0.92-1.00)	0.99 (0.96-1.00)
	Sierra Leone	0.97 (0.95-0.99)	0.94 (0.92-0.96)	0.96 (0.94-0.97)
	South Africa	0.98 (0.97-1.00)	0.94 (0.92-0.97)	0.96 (0.95-0.98)
	Overall	0.98 (0.98-0.99)	0.97 (0.96-0.98)	0.98 (0.97-0.98)
Location	Bangladesh	0.95 (0.93-0.98)	1.00 (0.99-1.00)	0.97 (0.96-0.99)
	Ethiopia	0.93 (0.88-0.97)	0.86 (0.80-0.91)	0.89 (0.85-0.93)
	Kenya	0.89 (0.84-0.94)	0.93 (0.89-0.97)	0.91 (0.87-0.94)
	Mozambique	1.00 (1.00-1.00)	1.00 (1.00-1.00)	1.00 (1.00-1.00)
	Nigeria	0.72 (0.58-0.85)	1.00 (1.00-1.00)	0.84 (0.72-0.92)
	Sierra Leone	0.76 (0.69-0.83)	0.98 (0.95-1.00)	0.86 (0.80-0.90)
	South Africa	0.96 (0.91-1.00)	0.81 (0.74-0.89)	0.88 (0.83-0.92)
	Overall	0.90 (0.88-0.92)	0.93 (0.92-0.95)	0.91 (0.90-0.93)
Name	Bangladesh	0.93 (0.86-1.00)	0.98 (0.92-1.00)	0.96 (0.91-0.99)
	Ethiopia	0.80 (0.61-0.96)	1.00 (1.00-1.00)	0.89 (0.76-0.98)
	Kenya	1.00 (1.00-1.00)	0.83 (0.70-0.94)	0.91 (0.82-0.97)
	Sierra Leone	1.00 (1.00-1.00)	0.88 (0.75-0.97)	0.94 (0.86-0.99)
	South Africa	1.00 (1.00-1.00)	1.00 (1.00-1.00)	1.00 (1.00-1.00)
	Overall	0.93 (0.90-0.96)	0.94 (0.91-0.97)	0.94 (0.91-0.96)
Pediatric Clinical Abstractions (PCA)				
Date	Ethiopia	0.91 (0.76-1.00)	1.00 (1.00-1.00)	0.95 (0.87-1.00)
	Kenya	0.98 (0.97-0.99)	0.99 (0.98-0.99)	0.98 (0.98-0.99)
	Mozambique	1.00 (1.00-1.00)	0.94 (0.87-1.00)	0.97 (0.93-1.00)
	Nigeria	1.00 (1.00-1.00)	0.97 (0.90-1.00)	0.98 (0.94-1.00)

Continued on next page

Table S6 – continued from previous page

Entity	Country	Precision (95% CI)	Recall (95% CI)	F_1-score (95% CI)
	Sierra Leone	0.99 (0.98-1.00)	0.93 (0.90-0.95)	0.96 (0.94-0.97)
	South Africa	0.98 (0.96-0.99)	0.96 (0.94-0.98)	0.97 (0.96-0.98)
	Overall	0.98 (0.98-0.99)	0.97 (0.96-0.98)	0.97 (0.97-0.98)
Location	Ethiopia	0.90 (0.83-0.96)	1.00 (1.00-1.00)	0.95 (0.91-0.98)
	Kenya	0.89 (0.84-0.93)	0.97 (0.94-0.99)	0.92 (0.90-0.95)
	Mozambique	0.87 (0.79-0.95)	1.00 (1.00-1.00)	0.93 (0.88-0.97)
	Nigeria	0.71 (0.55-0.85)	1.00 (1.00-1.00)	0.83 (0.70-0.92)
	Sierra Leone	0.95 (0.92-0.97)	0.98 (0.96-1.00)	0.96 (0.94-0.98)
	South Africa	0.89 (0.84-0.94)	0.88 (0.84-0.93)	0.89 (0.85-0.92)
	Overall	0.90 (0.88-0.92)	0.96 (0.94-0.97)	0.93 (0.92-0.94)
Name	Ethiopia	0.83 (0.65-0.96)	1.00 (1.00-1.00)	0.91 (0.80-0.98)
	Kenya	0.87 (0.78-0.95)	0.91 (0.83-0.98)	0.89 (0.82-0.94)
	Sierra Leone	0.74 (0.66-0.80)	0.98 (0.96-1.00)	0.84 (0.79-0.89)
	South Africa	0.69 (0.50-0.88)	1.00 (1.00-1.00)	0.82 (0.67-0.92)
	Overall	0.76 (0.71-0.81)	0.97 (0.94-0.99)	0.85 (0.82-0.88)

S5. Single Module Performance

As detailed in Table S7, for the enity "name", the Full System consistently demonstrated superior efficacy, achieving the highest F_1-scores across all datasets (0.75–0.94). By integrating the high sensitivity of the PII detection agent (Recall: 0.94–0.98) with the strict specificity of the rule-based engine, the Full System successfully mitigated the trade-off between over-redaction and data leakage. Consequently, it significantly outperformed both the standalone sub-modules and the Piiredact baseline, ensuring robust privacy preservation while minimizing the loss of clinical context.

Dataset	Method	Precision (95% CI)	Recall (95% CI)	F_1-score (95% CI)
VA	Full System	0.61 (0.55-0.67)	0.98 (0.96-1.00)	0.75 (0.70-0.80)
	PII detection agent	0.45 (0.40-0.51)	0.94 (0.90-0.98)	0.61 (0.56-0.66)
	Piiredact	0.32 (0.28-0.37)	0.85 (0.80-0.91)	0.47 (0.41-0.52)
	High-precision rule-based engine	1.00 (1.00-1.00)	0.26 (0.20-0.33)	0.41 (0.33-0.50)
MCA	Full System	0.93 (0.90-0.96)	0.94 (0.91-0.97)	0.94 (0.91-0.96)
	PII detection agent	0.58 (0.52-0.63)	0.96 (0.93-0.98)	0.72 (0.68-0.76)
	Piiredact	0.28 (0.24-0.32)	0.66 (0.59-0.73)	0.39 (0.34-0.44)
	High-precision rule-based engine	0.96 (0.92-1.00)	0.42 (0.36-0.49)	0.58 (0.51-0.65)
PCA	Full System	0.76 (0.71-0.81)	0.97 (0.94-0.99)	0.85 (0.82-0.88)
	PII detection agent	0.61 (0.55-0.66)	0.95 (0.92-0.98)	0.74 (0.70-0.78)
	Piiredact	0.46 (0.41-0.52)	0.69 (0.63-0.76)	0.56 (0.50-0.61)
	High-precision rule-based engine	0.89 (0.80-0.96)	0.24 (0.18-0.30)	0.38 (0.30-0.45)

Table S7: Module-by-Module Performance Metrics for Named Entity Recognition.

For the entity "location", Table S8 demonstrates the robust capability of the Full System in identifying location entities, yielding consistent F_1-scores between 0.91 and 0.94 across all datasets. The system exhibited a significant performance leap over the Piiredact baseline, which failed to identify the vast majority of location data (Recall: 0.11–0.24). Furthermore, while the individual components (the PII detection agent and the rule-based engine) struggled with precision (0.49–0.63), the Full System effectively combined these inputs. It maintained a stable precision of 0.90 while achieving exceptional recall (0.93–0.99), ensuring that geographical identifiers were removed with high reliability and minimal over-redaction.

Dataset	Method	Precision (95% CI)	Recall (95% CI)	F_1-score (95% CI)
VA	Full System	0.90 (0.88-0.91)	0.99 (0.98-0.99)	0.94 (0.93-0.95)
	PII detection agent	0.49 (0.46-0.51)	0.88 (0.86-0.90)	0.63 (0.61-0.65)
	Piiredact	0.64 (0.59-0.69)	0.24 (0.21-0.26)	0.35 (0.31-0.38)
	High-precision rule-based engine	0.54 (0.52-0.57)	0.90 (0.89-0.92)	0.68 (0.66-0.70)
MCA	Full System	0.90 (0.88-0.92)	0.93 (0.92-0.95)	0.91 (0.90-0.93)
	PII detection agent	0.54 (0.51-0.57)	0.79 (0.76-0.81)	0.64 (0.62-0.66)
	Piiredact	0.41 (0.36-0.45)	0.16 (0.14-0.18)	0.23 (0.20-0.26)
	High-precision rule-based engine	0.61 (0.58-0.64)	0.85 (0.82-0.87)	0.71 (0.69-0.73)
PCA	Full System	0.90 (0.88-0.92)	0.96 (0.94-0.97)	0.93 (0.92-0.94)
	PII detection agent	0.55 (0.53-0.58)	0.87 (0.84-0.89)	0.68 (0.65-0.70)
	Piiredact	0.44 (0.37-0.51)	0.11 (0.09-0.13)	0.17 (0.14-0.20)
	High-precision rule-based engine	0.63 (0.60-0.66)	0.77 (0.74-0.80)	0.69 (0.67-0.72)

Table S8: Performance Metrics for Location Entity Recognition on the annotated data.

As presented in Table S9, the extraction of temporal entities yielded the highest overall performance metrics among all PHI categories. The Full System achieved near-perfect fidelity, with F_1-scores ranging from 0.95 to 0.98 and precision reaching 1.00 (95% CI: 0.99-1.00) in the VA cohort. Notably, unlike name or location entities, the high-precision rule-based engine demonstrated exceptional standalone performance (Recall: 0.90–0.97), reflecting the highly structured syntactic patterns of dates in clinical text. However, the Full System still offered a critical advantage by combining this rule-based precision with the high sensitivity of the PII detection agent (Recall: 0.96–0.99), effectively eliminating the substantial over-redaction observed in the standalone agent (Precision: 0.60–0.63) while far surpassing the low recall of the Piiredact baseline of 0.34 (95% CI: 0.31-0.37).

Dataset	Method	Precision (95% CI)	Recall (95% CI)	F_1-score (95% CI)
VA	Full System	1.00 (0.99-1.00)	0.91 (0.89-0.93)	0.95 (0.94-0.96)
	PII detection agent	0.62 (0.59-0.65)	0.98 (0.97-0.99)	0.76 (0.74-0.78)
	Piiredact	0.48 (0.44-0.52)	0.34 (0.31-0.37)	0.40 (0.36-0.43)
	High-precision rule-based engine	0.87 (0.84-0.89)	0.90 (0.88-0.92)	0.88 (0.87-0.90)
MCA	Full System	0.98 (0.98-0.99)	0.97 (0.96-0.98)	0.98 (0.97-0.98)
	PII detection agent	0.63 (0.62-0.65)	0.99 (0.98-0.99)	0.77 (0.76-0.78)
	Piiredact	0.50 (0.46-0.53)	0.20 (0.18-0.21)	0.28 (0.26-0.30)
	High-precision rule-based engine	0.97 (0.96-0.97)	0.97 (0.97-0.98)	0.97 (0.97-0.98)
PCA	Full System	0.98 (0.98-0.99)	0.97 (0.96-0.98)	0.97 (0.97-0.98)
	PII detection agent	0.60 (0.58-0.62)	0.96 (0.95-0.97)	0.74 (0.72-0.75)
	Piiredact	0.42 (0.38-0.46)	0.13 (0.12-0.15)	0.20 (0.18-0.23)
	High-precision rule-based engine	0.97 (0.96-0.98)	0.97 (0.96-0.98)	0.97 (0.96-0.97)

Table S9: Performance Metrics for Date Entity Recognition on the annotated data.