# Evaluating Open-Source Models for De-Identification of Unstructured Clinical Text

Singapore Healthcare Management 2024

Singapore General Hospital
SingHealth

**Chia Sing Yi, Keith Tan Yan Yang**
Health Services Research Unit
Singapore General Hospital

## Introduction

Free-text clinical notes contain rich information about patient health and clinical care details, but it is broadly left unused due to the presence of personally identifiable information (PII) to protect patient confidentiality.

Traditional methods for using unstructured text includes:
- Manual review and masking of PII, or
- Setting up restricted access sandbox environments

These methods are cumbersome and impractical for large-scale research and collaborations.

## Aim

Evaluate the potential of utilising open-source models to automate de-identification of unstructured local clinical text.

## Methodology

We manually annotated **5,869** clinical notes from 4 domains:

| Text Domains | Count |
|---|---|
| Clindoc Notes | 248 |
| Lab General Comments | 721 |
| SCM Order Comments | 4,500 |
| Radiology Reports | 400 |

Common PIIs: Names, IDs, MRNs, Addresses, and Phone numbers

Two open-source models, Protected Health Information filter (Philter) and Stanford AIMI, were selected for testing and evaluation based on their published performance scores. The model performance was assessed based on their Precision, Recall, and F2 Score.

- Recall = TP / (TP + FN)
  – High recall → FN is low → minimise missed PHI
- Precision = TP / (TP + FP)
  – High precision → FP is low → minimise information lost
- F2 Score = (5* Precision * Recall)/ ((4 * Precision) + Recall)

| | **Actual Positive** | **Actual Negative** |
|---|---|---|
| **Predicted Positive** | True Positive (TP) *correctly detects PHI* | False Positive (FP) *mistaken non-PHI as PHI* |
| **Predicted Negative** | False Negative (FN) *couldn't detect PHI* | True Negative (TN) *correctly classify non-PHI* |

## Conclusion

The Stanford model showed promising results for large-scale de-identification of unstructured text with its ability to balance accurate redaction of PII and preserving the integrity of the text. The findings served as a baseline for model comparisons and a motivation to formalise a governance framework for acceptable risk thresholds.

## Results

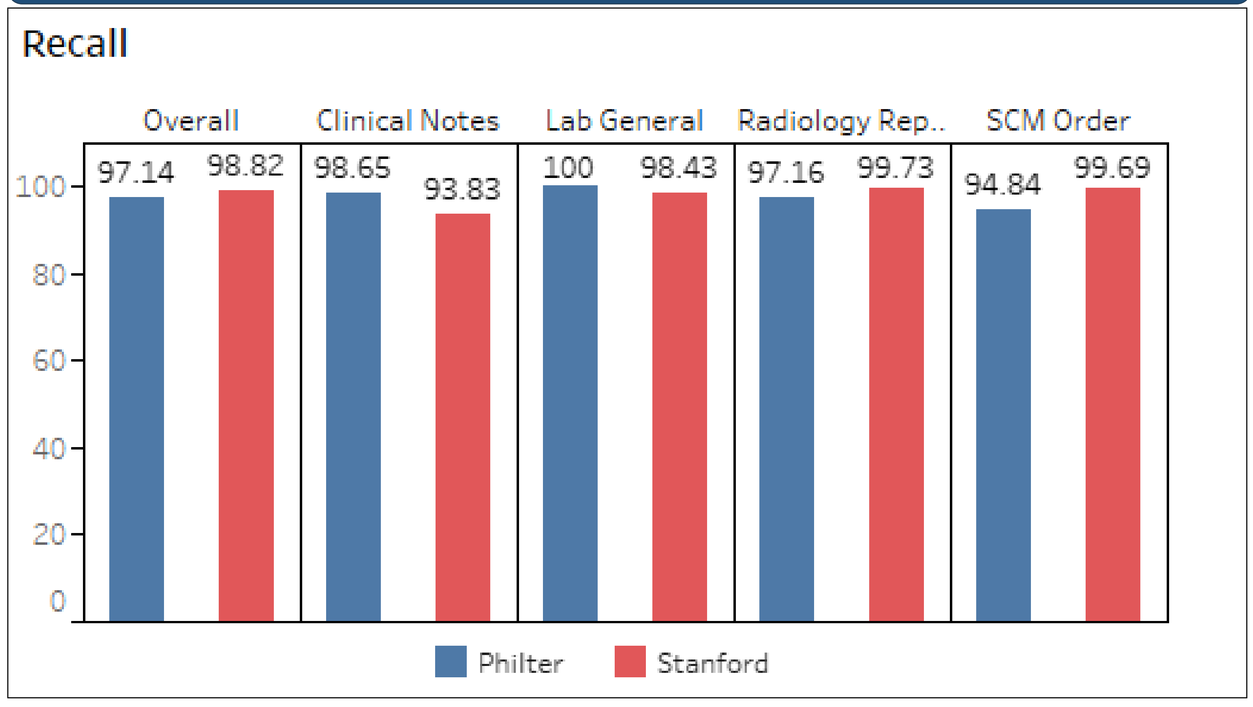


Figure 1. Recall of Philter and Stanford models across different domains

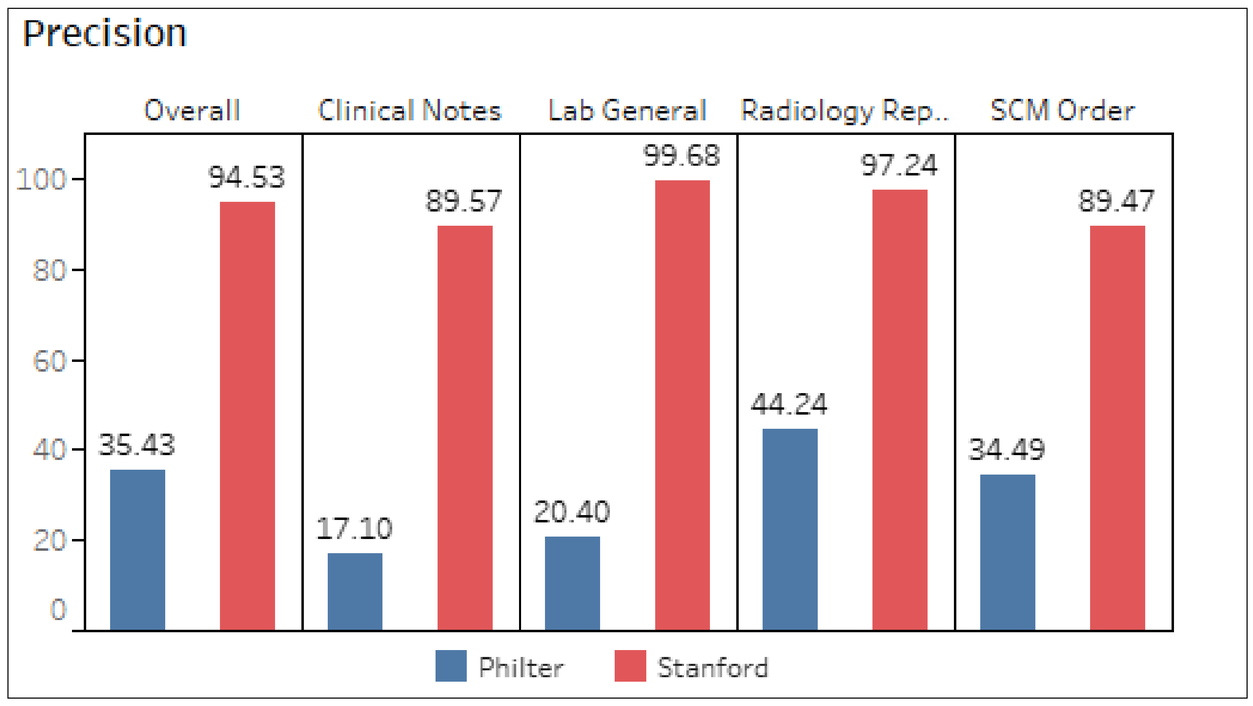


Figure 2. Precision of Philter and Stanford models across different domains

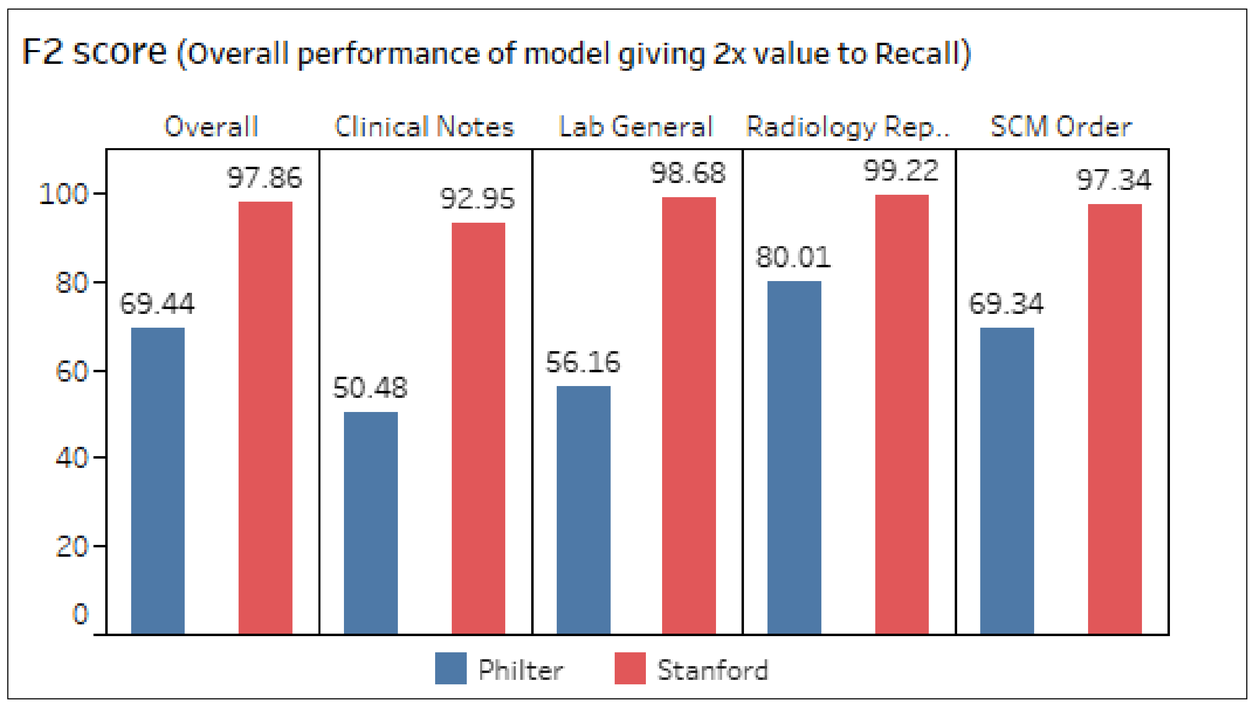


Figure 3. Overall performance of model ; F2 Score = (5* Precision * Recall)/ ((4 * Precision) + Recall)

Overall, both models produced high recall rates – Stanford 98.8% and Philter 97.1%. However, the Philter model had a low precision rate of 39.3% which results in over-masking, hence significantly reduced the utility of the de-identified text. Conversely, the Stanford model achieved a precision of 94.5%, effectively removing identifiers while preserving the integrity and usefulness of the de-identified text. Notably, both models faced challenges with recognising Chinese names and abbreviated initials which resulted in occasions of partial masking.