NATIONAL CENTER FOR HEALTH STATISTICS

What's in a Name Field?

Frances McCarty[1], Ben Rogers[2], Jessie Parker[1], Cordell Golden[1] National Center for Health Statistics, [1]Division of Analysis and Epidemiology, Data Linkage Methodology and Analysis Branch [2]Divsion of Research Methodology, Office of the Director

Overview

Background

- Data linkage accuracy depends on the quality of the data fields.
- Data review and cleaning are essential to address data quality and improve linkage accuracy.
- Automating the data review and cleaning process reduces time-consuming manual review.

Challenges

- Type of data field (date vs name) affects the level of effort needed for review.
- Name fields may contain **non-name text** that needs to be identified and removed.

Objective

- Examine the use of artificial intelligence (AI) based large language models (LLM) and simple rule-based approaches **to identify non-name text in name fields**.

Methods

Data

- Testing data created with R (v4.1.3 (2022-03-10)) package randomNames.
- Names randomly generated from public data based on sampling gender and ethnicity.
- Name list includes records with valid first and last names (n=9,949).
- Name list supplemented with **non-name text** ("pilot study", "department funded") that might be in survey data/administrative records (n=166).
- Non-name text could appear in first or last name or both
  - » Indicates records that should be flagged for review.

Identification of non-name records

Large Language Models (LLM) with few-shot prompting

- Only applied to last name
  - » Determine how well LLM could do if data leakage was only in one field.
- GPT-3.5 fined tuned from GPT-3 using a process named Reinforcement Learning from Human Feedback to provide better results following instructions.
- Few-shot prompting uses the inherent knowledge present within LLMs adapted to the specific task using in-context learning.
  - » Provide a description of the task at hand, such as "The following are records of names from a form with a data contamination causing other form fields to be saved under the name section. Please classify the following text as an entry in the name category by responding with a 1 or 0. The following are examples:"

AI Chatbot (cdc.gov)

- Only applied to last name
  - » Determine how well LLM could do if data leakage was only in one field.
- Uses GPT-3.5, with training data up to September 2021.
- Name list provided and the following question posed: "Can you identify the words in the list that are not last names?"

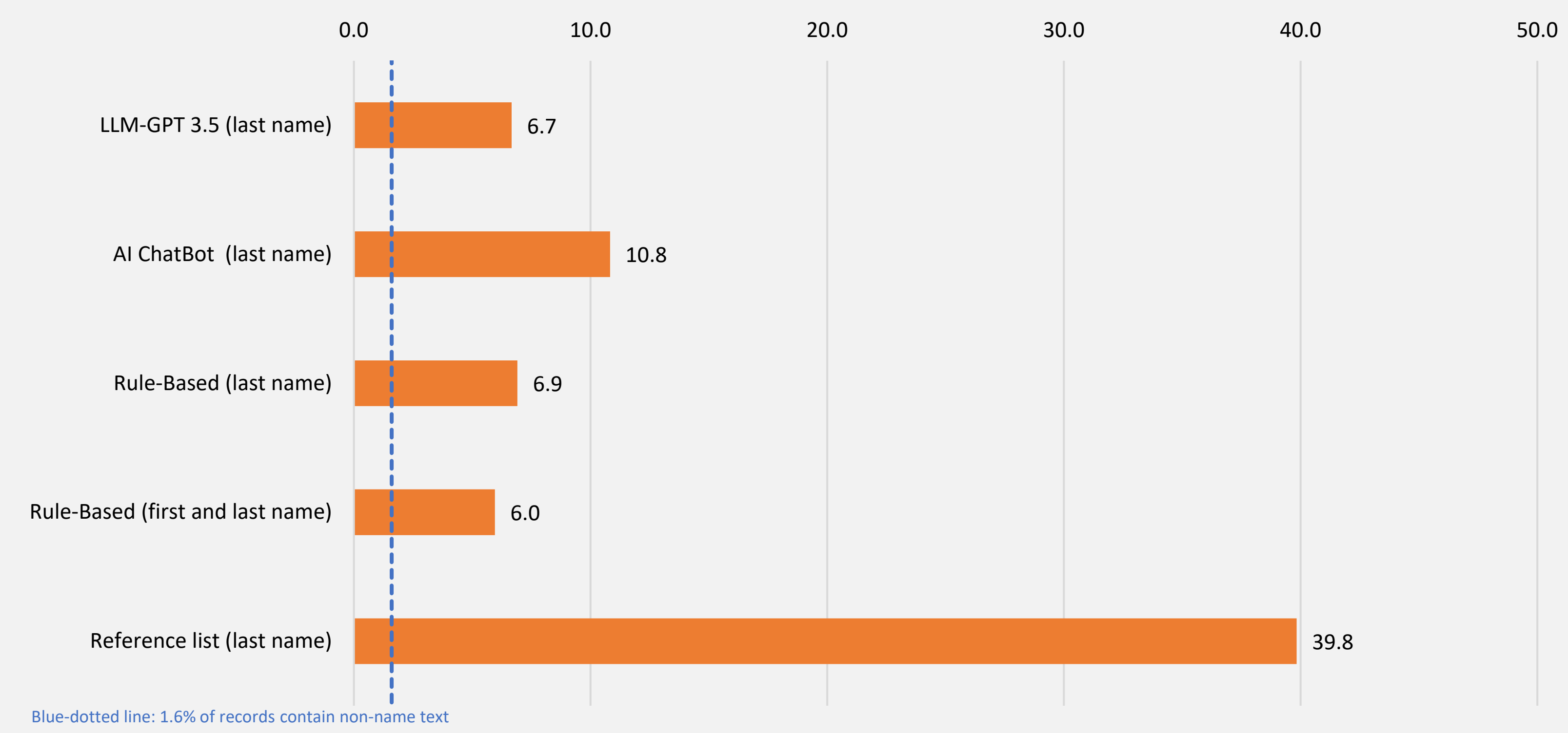
Rule-based

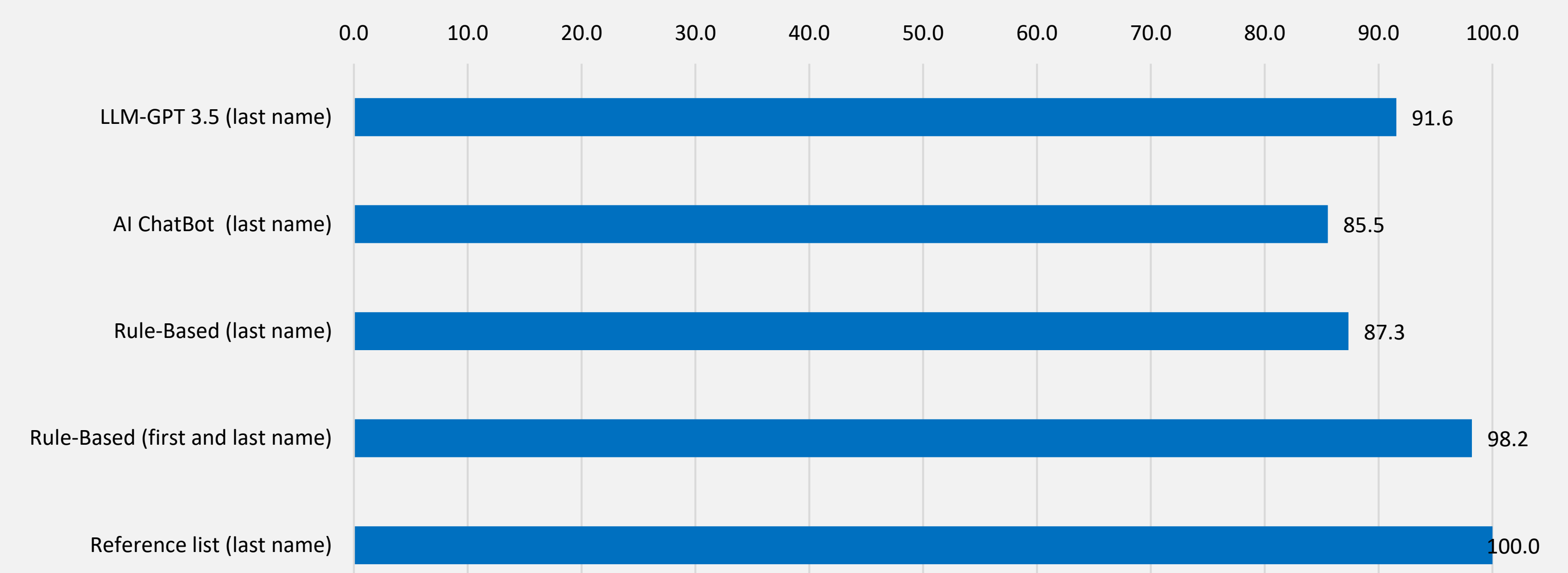
- Applied to both first and last name
- Uses name features – number of words (w) and characters (c)
  - » Last name only: (w>=2 & average c/w<=5) OR w>2 OR total c>11
  - » First and last name: (w>=3 & average c/w<=5) OR w>4 OR total c>19

Comparison with reference list of "valid" last names

Results

Number of records = 10,115 (actual names, n=9,949 / non-name text, n=166)





Types of records that were not flagged when they should have been

- LLM-GPT 3.5 (last name)
  - » Last name field contained both first and last name text and the first name field contained non-name text
- AI Chatbot (last name)
  - » Last name field contained both first and last name text and the first name field contained non-name text
  - » Last name field contained the following text: code, trial a
- Rule-based (last name)
  - » Last name field contained the following text: code, program, radiology
- Rule-based (first and last name)
  - » Last name field contained the following text: code, program, radiology; First name field contains non-name text (two words) : study patient, center study, patient services
- Comparison with reference list of "valid" names (last names)
  - » All records with non-name text flagged
  - » High percentage of false positives (close to 40% of records would require review)

Conclusions

LLM correctly identified over 90% of records with non-name text

- Pros
  - » Potentially easy to automate
  - » Reasonable number of records flagged for review
- Cons
  - » Requires expertise in LLM
  - » Application may not be approved for use with actual PII
  - » Potential issues processing large data sets
  - » Prompts and method of accessing can impact AI performance

AI ChatBot (cdc.gov) correctly identified about 85% of records with non-name text

- Pros
  - » Potentially easy to use
- Cons
  - » Application may not be approved for use with actual PII
  - » Potential issues processing large data sets
  - » Prompts and method of accessing can impact AI performance

Rule-Based (last name) correctly identified about 87% of records with non-name text, while rule-Based (first and last name) correctly identified about 98% of records with non-name text

- Pros
  - » Easy to implement with code
- Cons
  - » Criteria for the flag might change depending on the application

Reference list (last name) correctly identified all records with non-name text, but almost 40% of all records were flagged for review

- Pros
  - » Easy to implement with statistical programming code
- Cons
  - » Heavily dependent on accuracy and quality of the list used
  - » Name lists are not exhaustive and not always readily accessible
  - » Likely to flag an excessive number of records

References

Chung, H.W., Hou, L., Longpre, S., Zoph, B., Tay, Y., Fedus, W., Li, Y., Wang, X., Dehghani, M., Brahma, S. and Webson, A. (2022). "Scaling instruction-finetuned language models". arXiv preprint arXiv:2210.11416.

Damian W. Betebenner (2021). randomNames: Function for Generating Random Names and a Dataset. (R package version 1.5-0.0 URL https://cran.r-project.org/package=randomNames

Open AI. (2022). "Introducing ChatGPT". https://openai.com/blog/chatgpt

R Core Team (2022). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

Contact Info

Frances McCarty, CDC/NCHS/DLMAB
FMcCarty@cdc.gov
301-458-4247