

Leveraging Open-Source Large Language Models for Automated Patient Record Linkage at Central Cancer Registries

Mohammad Beheshti, MSHI, Lovedeep Gondara, PhD, Mihail Popescu, PhD, Prasad Calyam, PhD, Iris Zachary, PhD



College of Health Sciences
University of Missouri

INTRODUCTION

What is the problem?

- Healthcare data is fragmented across multiple sources (hospitals, physician offices, labs).
- Central cancer registries serve as a prime example of healthcare organizations facing these challenges as they aggregate data from multiple sources including hospitals, physician offices, and pathology labs to track cancer incidence, treatment, and outcomes.
- Cancer Registries rely on data linkage for various purposes including identifying non-reported cancer cases and consolidating duplicate records.

Key Challenges:

- Lack of a universal unique patient identifier across healthcare providers.
- Missing and inconsistent patient information.
- Variability in data formats across different healthcare providers.
- Errors in patient data (e.g., typos, name changes, incorrect birth dates).

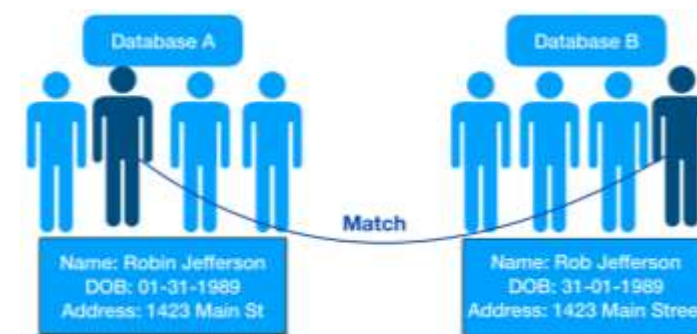
INTRODUCTION

What is Record Linkage?

Record linkage or entity resolution is the process of linking records from different data sources that refer to the same entity, such as a patient. There are two traditional approaches to record linkage:

1) Deterministic Linkage

- Relies on exact matches of specific identifiers (e.g., name, DOB, SSN).
- Fails when key identifiers (e.g., SSN, full name) are missing or incomplete.
- Minor typos or formatting differences can prevent accurate matching.



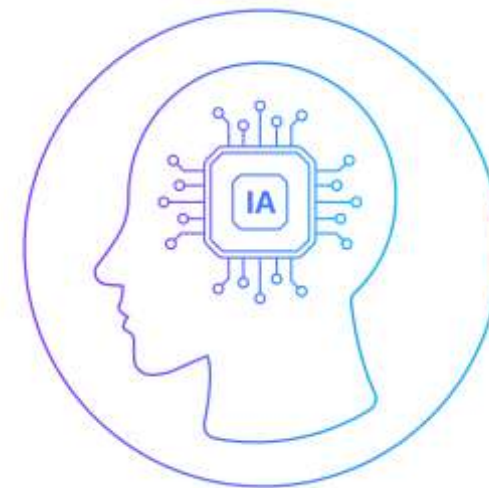
2) Probabilistic Linkage

- Uses statistical techniques to calculate the likelihood that two records refer to the same patient.
- Matching accuracy depends on setting an appropriate similarity threshold, which is dataset-specific.
- Often requires human validation to confirm uncertain matches, making the process time-consuming.
- The manual review process is repetitive and tedious.

INTRODUCTION

Language Models for Entity Resolution

- Transformer-based language models (e.g., BERT, RoBERTa, Sentence-BERT) have been used in the literature for general entity matching.
- Studies such as Li et al. (Ditto) demonstrate the ability of fine-tuned transformer models to improve entity resolution in retail and e-commerce datasets.
- More recently, LLMs (e.g., ChatGPT, Llama) have been explored for text-based matching tasks, showing strong performance.



What is the gap?

- Most studies focused on general entity resolution especially product matching.
- Use of benchmark datasets with synthetically created errors for evaluation which lack real-world complexities.
- Lack of studies with real-world patient data.

INTRODUCTION

Study Goals

1. Assess the feasibility of using language models as a blocking technique in patient record linkage and compare their performance with the traditional methods of blocking.
2. Investigate the use of language models for matching patient records and evaluate their performance.
3. Identify key challenges in applying language models to patient record linkage and propose strategies to enhance their reliability and scalability.



➤ Methods | Data Prep

Data Source

- This study utilized data from the Missouri Cancer Registry and Research Center.
- Two primary datasets selected for record linkage:
 - **Dataset_A:** Consolidated cancer patient records (CRS Plus).
 - **Dataset_B:** Pathology reports submitted electronically (eMaRC Plus).

Data Split

- **Train:** 2022 data was used for **fine-tuning** language models.
- **Test:** 2021 data was used for **testing** the models.



Methods | Data Prep

Preparing the Linked Data

- We performed a probabilistic record linkage using the Match*Pro software.
- The linkage configuration was executed for both train and test datasets independently.
- The resulting record pairs (Dataset_AB) were manually reviewed by experienced human annotators and labeled as Match or Non-Match based on their identifiers.

Field Name	Blocking Strategy	Matching Strategy
First Name	Soundex (Phonetic Matching)	Jaro-Winkler Distance = 0.8
Middle Name	-	Jaro-Winkler Distance = 0.8
Last Name	Soundex (Phonetic Matching)	Jaro-Winkler Distance = 0.8
Birth Date	Exact Match	Same Month and Year
Sex	-	Exact Match
Social Security Number	Exact Match	2 Edits or Transpositions

	Dataset_A	Dataset_B	Dataset_AB	Labels
Train (2022)	51,943	29,552	58,383	Non-Match: 54,858 Match: 3,525
Test (2021)	51,781	26,958	52,917	Non-Match: 50,561 Match: 2,356

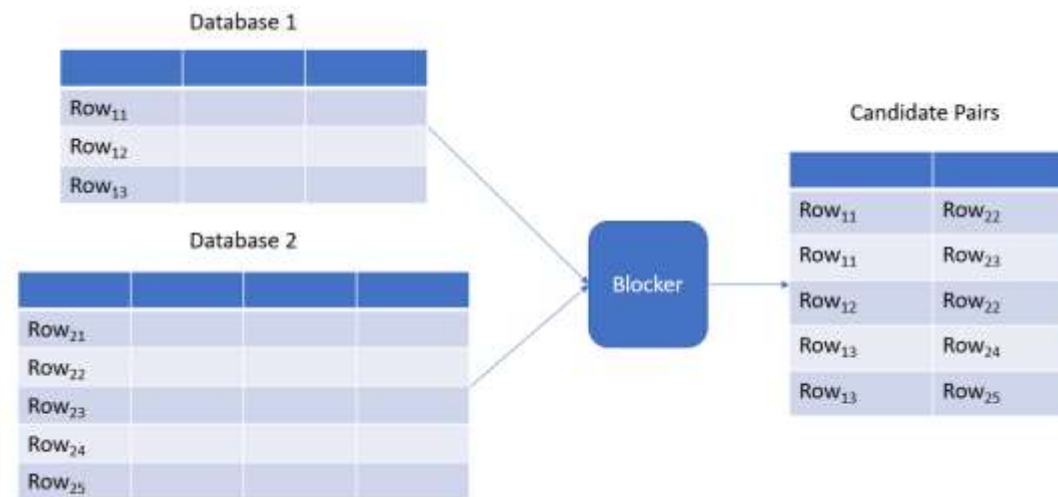
➤ Methods | Experiment 1: Blocking

What is Blocking?

- Reduces the number of candidate record pairs for matching.
- Without blocking, record comparisons scale exponentially.

Traditional Blocking Methods

- Exact Matching: Requires identical field values.
- Phonetic Encoding: Soundex, Metaphone
- Distance-Based: Jaro-Winkler, Levenshtein.
- Clustering Techniques: Canopy, Sorted Neighborhood.



Methods | Experiment 1: Blocking

Experiment Setup

- We used SentenceTransformers to fine-tune RoBERTa for semantic similarity embeddings.
- Fields with a high rate of missing data, including SSN and Address, were excluded from the blocking experiment.
- Each record pair in Dataset_AB was independently serialized into a string as follows:

Serialize (A) ::= [FirstName_A] [MiddleName_A] [LastName_A] [BirthDate_A] [Sex_A]

Serialize (B) ::= [FirstName_B] [MiddleName_B] [LastName_B] [BirthDate_B] [Sex_B]

- **Example:**

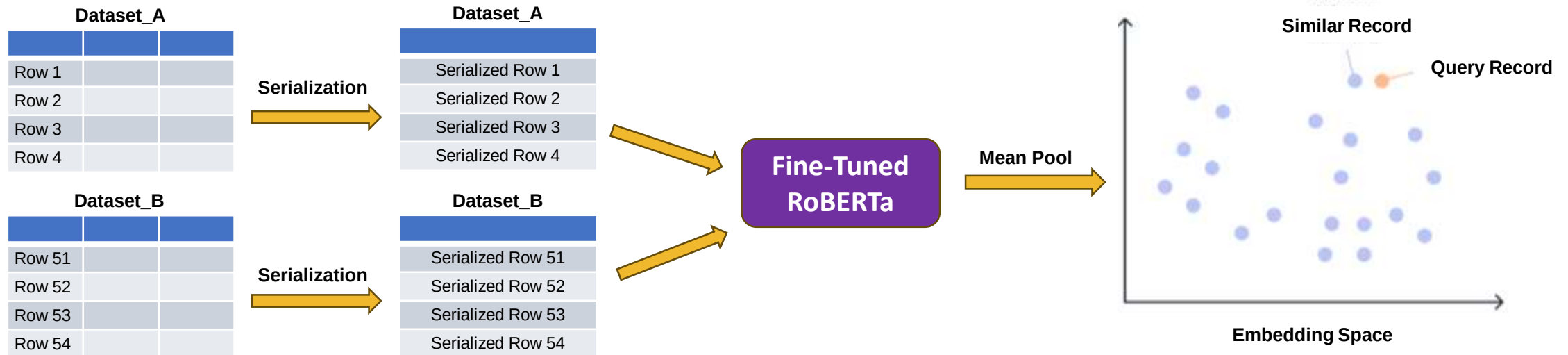
Jonathan M. Doe 19850715 Male

- Serialized string pairs were then fed into the model along with their corresponding labels (Overall Similarity Scores).
- We applied mean pooling and used Cosine Similarity Loss.
- Trained for 5 epochs with a batch size of 64, using the SBERT's default training arguments.

➤ Methods | Experiment 1: Blocking

Blocking Evaluation

- The fine-tuned model was utilized to generate embeddings for each serialized record in both test datasets using mean pooling.
- Record embeddings from Dataset_A were queried against those from Dataset_B using the K-Nearest Neighbors (KNN) algorithm implemented via the FAISS library.
- A fixed **K** was used along with an adjustable cosine similarity threshold to enhance the efficiency of the blocking.
- The resulting candidate record pairs were evaluated against the ground truth test set (Dataset_AB).



Methods | Experiment 2: Matching

Experiment Setup

- We fine-tuned RoBERTaForClassification, which employs RoBERTa as the base model with an additional linear layer, known as the classification head.
- We adopted the serialization method used in Ditto to structure the records for model input:

Serialize (A) ::= [COL] attr₁ [VAL] val₁ ... [COL] attr_k [VAL] val_k

Serialize (B) ::= [COL] attr₁ [VAL] val₁ ... [COL] attr_k [VAL] val_k

- The serialized records were then concatenated as a pair using the following format:

Serialize (A, B) ::= [CLS] serialize(A) [SEP] serialize(B) [SEP]

- **Example:**

[CLS] [COL] First Name [VAL] John [COL] Last Name [VAL] Doe **[SEP]** [COL] First Name [VAL] Jonathan [COL] Last Name [VAL] Doe

Methods | Experiment 2: Matching

Experiment Setup

- Additionally, we fine-tuned several open-source LLMs.
- We did not employ a few-shot approach, as our experiments indicated that model decisions were highly influenced by the limited examples provided.
- Due to the high inference time of reasoning models like DeepSeek-R1, we evaluated and compared the zero-shot models on a subset of the test data where $[0.65 < \text{Overall Similarity Score} < 1.0]$.

Model Name	Fine-Tuned	Zero-Shot
unsloth/Llama-3.2-3B-Instruct	✓	✗
unsloth/mistral-7b-instruct-v0.3	✓	✗
unsloth/Meta-Llama-3.1-8B-Instruct	✓	✓
unsloth/Llama-3.3-70B-Instruct-bnb-4bit	✗	✓
unsloth/DeepSeek-R1-Distill-Llama-70B-bnb-4bit	✗	✓

Methods | Experiment 2: Matching

Prompt Format

```
prompt = f"""

You are given two patient records. Your task is to determine whether they
belong to the same individual. Consider factors such as name similarity, date
of birth, and other identifying attributes. Only respond with "Yes" or "No".

Record 1:
- First Name: {row['record1 First Name']}
- Middle Name: {row['record1 Middle Name']}
- Last Name: {row['record1 Last Name']}
- Date of Birth: {row['record1 Date of Birth']}
- SSN: {row['record1 SSN']}
- Sex: {row['record1 Sex']}
- Address: {row['record1 Address']}

Record 2:
- First Name: {row['record2 First Name']}
- Middle Name: {row['record2 Middle Name']}
- Last Name: {row['record2 Last Name']}
- Date of Birth: {row['record2 Date of Birth']}
- SSN: {row['record2 SSN']}
- Sex: {row['record2 Sex']}
- Address: {row['record2 Address']}

"""
```

> Methods | Experiment 2: Matching

Experiment Setup

- All the fine-tuned models were trained using LoRA with the following parameters:

Parameter	Epochs	Batch Size	Learning Rate	Optimizer	LR Scheduler	Warmup Ratio	Weight Decay	LoRA Rank	LoRA Alpha
Value	3	32	$2e^{-5}$	AdamW	Cosine	0.1	0.01	32	32

- The Unsloth library was employed to enhance computational efficiency and optimize VRAM consumption, enabling scalable fine-tuning and inference of LLMs.



Methods | Experiment 2: Matching

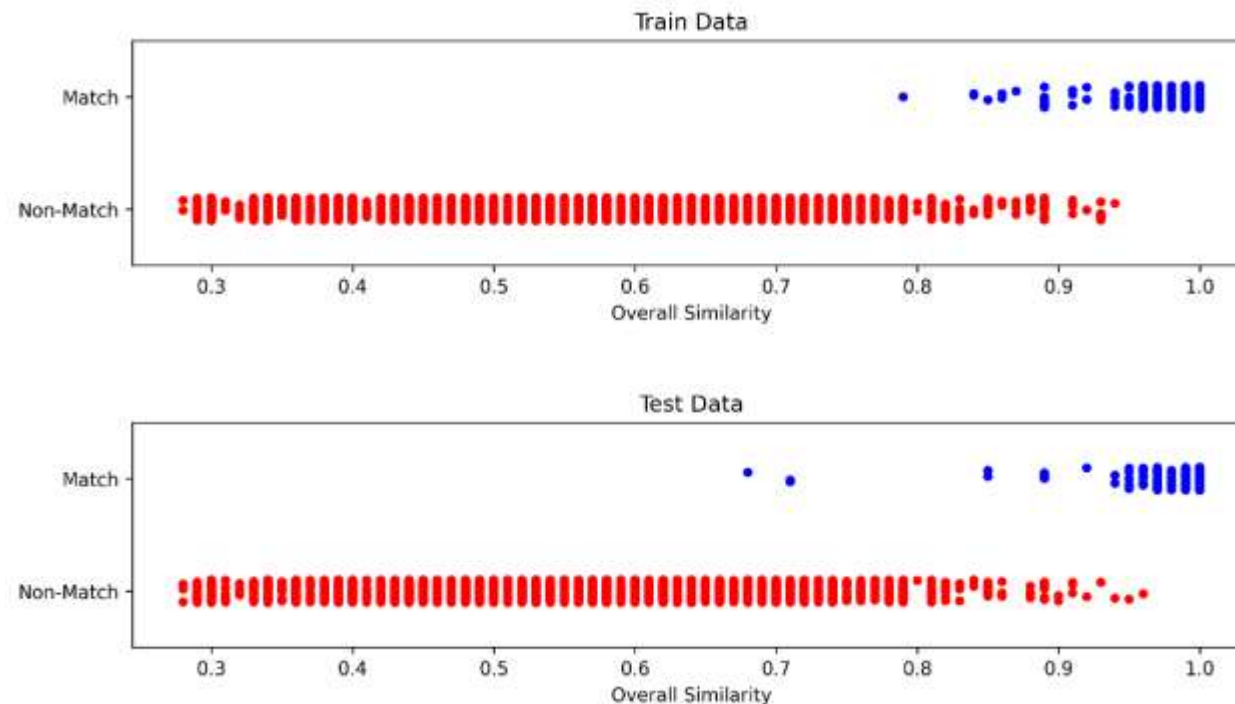
Matching Evaluation

- Models were used for inference on the test record pairs (Dataset_AB).
- The do_sample parameter was set to False during inference, enforcing a deterministic approach (Except DeepSeek-R1).
- The system prompt "You are a helpful assistant" was applied across both the fine-tuning and inference stages for all models except DeepSeek-R1.
- The model outputs were then compared against the ground truth labels to assess performance.
- All the experiments in this study were performed locally on a workstation with the following configuration:
 - **GPU:** 1 × Nvidia RTX A6000 (48GB VRAM)
 - **CPU:** AMD Ryzen Threadripper PRO 5955WX (16 Cores)
 - **RAM:** 128 GB

RESULTS

Record Pairs Distribution

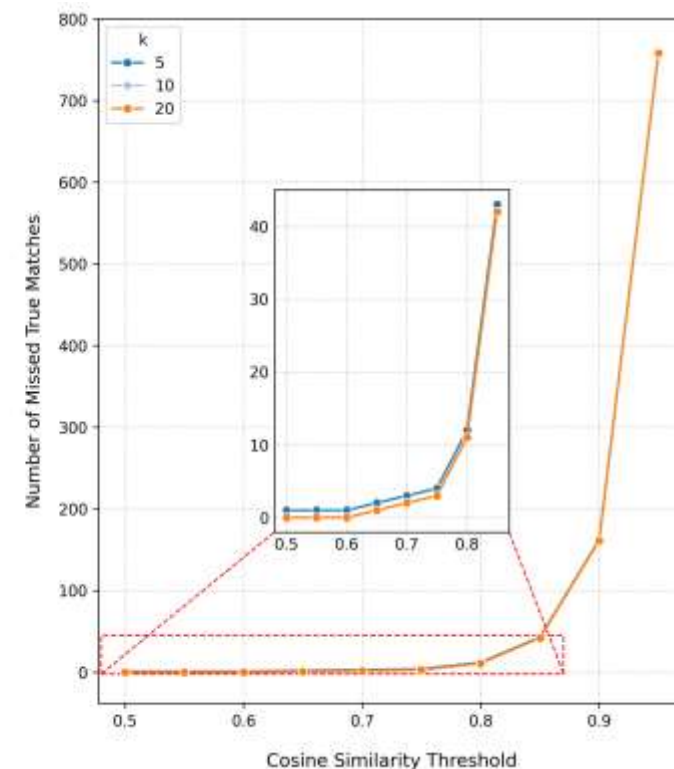
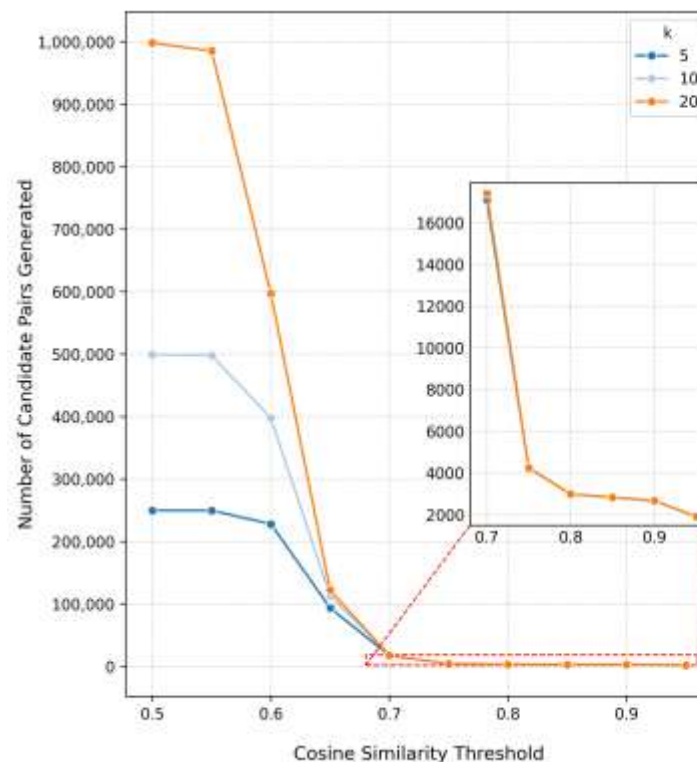
- Matching record pairs cluster above 0.8, with most near 0.9+, while non-matching pairs spread across lower similarity ranges.
- Overlap between matches and non-matches in the 0.85–0.95 range leads to potential false positives and false negatives.



RESULTS

Blocking Performance

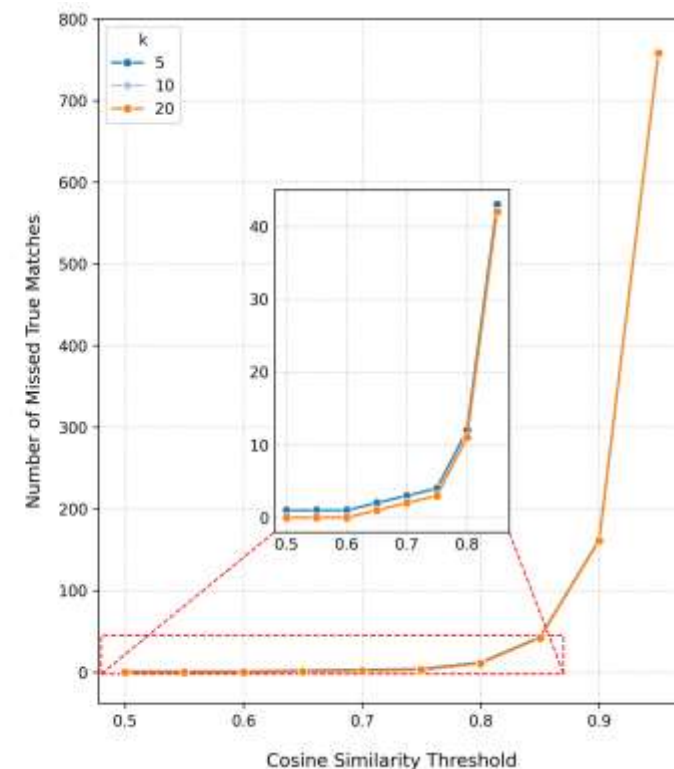
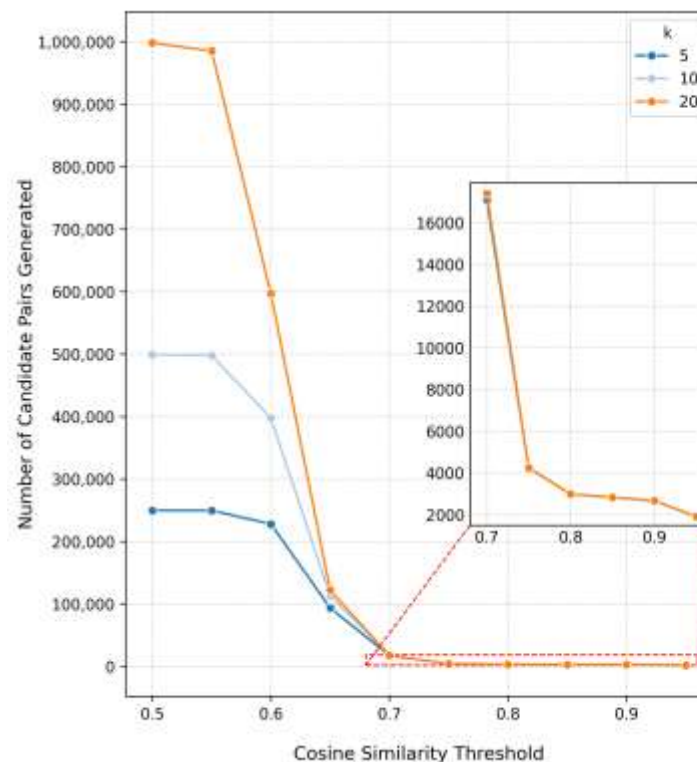
- The similarity threshold significantly affects recall and retrieval efficiency.
- Lower similarity thresholds (0.5–0.6) ensured nearly perfect recall but generated excessive candidate pairs (e.g., 500,000 pairs at 0.5 with K=10).
- Increasing the threshold to 0.7 significantly reduced the number of generated pairs to 17,396 resulting in two missed true matches.



RESULTS

Blocking Performance

- At a threshold of 0.75, retrieved pairs dropped to 4,250, with a loss of 3 true matches.
- At higher thresholds (0.8+), efficiency improved significantly, but recall dropped sharply.
- Increasing K from 5 to 10 slightly improved recall, but further increases had no additional impact.
- An optimal balance was found at K=10 and a similarity threshold of 0.75, maintaining high recall while reducing candidate pairs for efficient matching.



RESULTS

Matching Performance

- The fine-tuned Mistral-7B had the lowest number of incorrect predictions (FP + FN = 6).
- RoBERTa had the highest error count among the fine-tuned models (FP + FN = 27).
- Mistral-Small-24B had the best zero-shot performance.

Model	Fine-Tuned				Zero-Shot			
	FP	FN	FP + FN	F1 Score	FP	FN	FP + FN	F1 Score
RoBERTa	8	19	27	0.996995	—	—	—	—
Llama-3.2-3B	1	8	9	0.998999	—	—	—	—
Mistral-7B	0	6	6	0.999333	2,450	4	2,454	0.816149
Llama-3.1-8B	3	5	8	0.999111	0	839	839	0.887572
Mistral-Small-24B	0	11	11	0.998776	0	55	55	0.993823
Llama-3.3-70B	—	—	—	—	14	65	79	0.991135

RESULTS

DeepSeek-R1 VS Llama-3.3-70B

- DeepSeek-R1 was tested only in the zero-shot setting on a smaller subset of the test data (n = 2,736).

Model	Zero-Shot			
	FP	FN	FP + FN	F1 Score
DeepSeek-R1-70B	66	32	98	0.855
Llama-3.3-70B	7	43	50	0.909

- DeepSeek-R1 took approximately 26 hours to process the subset, whereas Llama-3.3-70B completed the same subset in just 30 minutes.



RESULTS

You are given two patient records. Your task is to determine whether they belong to the same individual. Consider factors such as name similarity, date of birth, and other identifying attributes. Only respond with "Yes" or "No".

Record 1:

First Name: BOBBY

Middle Name: M

Last Name: Smith

Date of Birth: 19450524

SSN: 698204485

Sex: 1

Address: 1234 Health Ave Apt 101, MO 65203

Record 2:

First Name: ROBERT

Middle Name: Unknown

Last Name: Smith

Date of Birth: 19450524

SSN: Unknown

Sex: 1

Address: Unknown

Okay, so I need to figure out if these two patient records belong to the same person. Let me break it down step by step.

First, looking at the first names: Record 1 has "BOBBY" and Record 2 has "ROBERT." Hmm, Bobby is a common nickname for Robert, right? So that's a point in favor of them being the same person because nicknames can vary but refer to the same name.

Next, the middle name. Record 1 shows "M," while Record 2 just says "Unknown." Since one has a middle initial and the other doesn't provide it, I can't really compare that. It's possible they might match if we knew both, but for now, it's not a contradiction.

The last names are both "Smith" in both records. That's a common surname, so while it's a point towards them being the same, it's not conclusive on its own because many people share that last name.

Now, the date of birth is exactly the same: 19450524. That's May 24, 1945. Dates of birth are pretty unique identifiers, especially when combined with other factors. So this is a strong indicator they could be the same person.

Looking at the SSN, Record 1 has 698204485, but Record 2 lists it as "Unknown." Without knowing Record 2's SSN, I can't confirm if they match. However, the DOB and name similarity are already aligning well.

Sex is marked as '1' in both records. Assuming that '1' stands for male, this matches. So that's consistent between them.

The address in Record 1 is specific: 1234 Health Ave Apt 101, MO 65203. Record 2 just says "Unknown." Since addresses can change over time and aren't always up-to-date in records, the fact that one has it and the other doesn't isn't necessarily a red flag. It could just mean that the information wasn't available or was outdated.

Putting it all together: The first name is a nickname of the other's full name, same last name, same DOB, same sex, but missing middle name, SSN, and address in Record 2. Given that DOB is a strong identifier and the names are likely related, I think these records point to the same person. The missing info in Record 2 doesn't necessarily mean they're different; it could just be incomplete data.

CONCLUSION

RoBERTa for Record Blocking

Pros

- Enabled flexible and scalable record blocking without the need for manual rule crafting.
- Generated well-balanced candidate pairs with high recall on unseen dataset.
- Reduced the number of unnecessary comparisons, improving efficiency.

- Rule-based: **52,917** Record Pairs

- RoBERTa: **4,250** Record Pairs



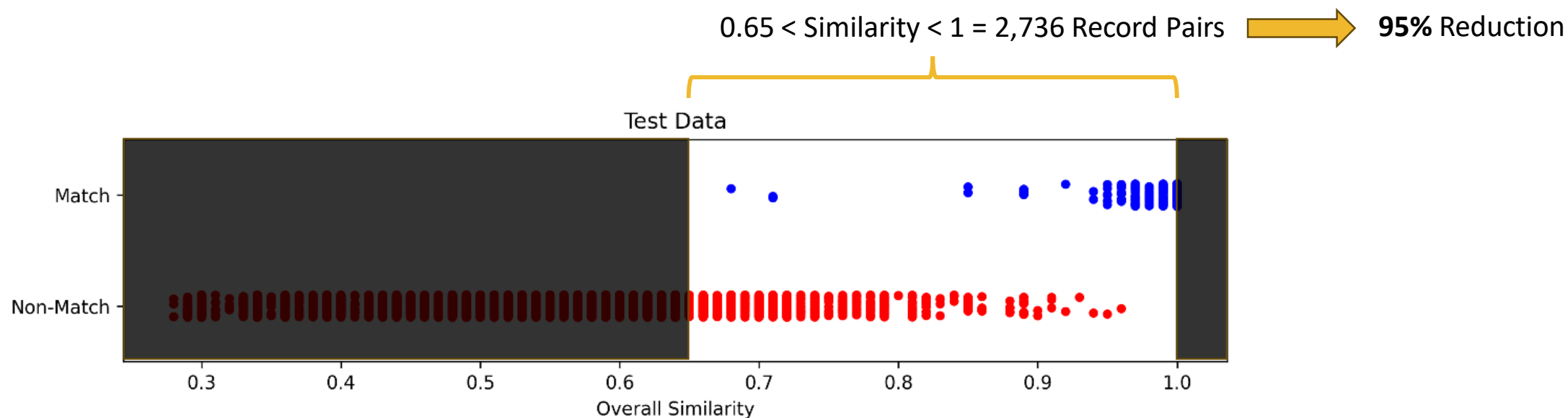
92% Reduction

> CONCLUSION

RoBERTa for Record Blocking

Cons

- Sacrifices a few matching records.
- Rule-based approach can achieve a better efficiency if combined with probabilistic overall similarity scores.



CONCLUSION

Language Models for Record Matching

- All fine-tuned models achieved near-perfect F1 score with minimal errors.
- All fine-tuned generative LLMs outperformed the baseline probabilistic matching (FP + FN = 17).
- Mistral-7B showed the best performance but **not** significantly better than Llama-3.1-8B or Llama-3.2-3B.
- Instruction finetuning yielded superior performance compared to conventional classification models such as RoBERTa.
- Mistral-Small-24B demonstrated great performance in a zero-shot setting; however, it underperformed compared to fine-tuned, smaller models such as Llama3.2-3B and even RoBERTa.
- While reasoning models such as DeepSeek-R1 offer a robust chain-of-thought process, their substantial computational time and resource requirements makes them less suitable for record linkage task compared to non-reasoning models.

FUTURE WORK

- Explore a Human-in-the-loop (HITL) approach.
- Utilize language models for tumor-level linkage addressing cases with multiple tumor instances.
- Implement the end-to-end linkage pipeline at MCR.



ACKNOWLEDGMENTS

Funding

- The Missouri Cancer Registry and Research Center (MCR) is supported in part by a collaborative agreement between the Centers for Disease Control and Prevention (CDC) and the Missouri Department of Health and Senior Services (DHSS) (NU5 8DP007130-02) and a Surveillance Contract between DHSS and the University of Missouri.
- Additionally, this study was funded through Data Modernization Initiative (DMI) under the Public Health Infrastructure Grant (PHIG).



Thank You!

Q&A

Mohammad Beheshti, MSHI

Informatics & Data Science Specialist

Missouri Cancer Registry and Research Center



College of Health Sciences
University of Missouri

Email: mbwnh@missouri.edu