

# LOINC Laboratory Diversifier: A Benchmark Dataset for Interoperability

Elisa Castagnari<sup>1</sup>, Ole Eigenbrod<sup>2</sup>, Honghan Wu<sup>3</sup>, and T.Ian Simpson<sup>1</sup>,

1. School of Informatics, University of Edinburgh, Edinburgh, UK. 2. Roche Diagnostics GmbH, Germany 3. School of Health and Wellbeing, University of Glasgow, Glasgow, UK.

## 1. Motivation

**Interoperability:** ability of different information systems, software applications, and data sources to exchange data and to use the information that has been exchanged in a reliable, meaningful, and efficient way.

**Goal:** Develop and evaluate methods that use large language models (LLMs) to generate high-quality, clinically meaningful synonyms for biomedical concepts to improve terminology mapping and interoperability between heterogeneous clinical and research data sources.

Concretely, the project aims to:

- Design and implement pipelines for generating synonyms for laboratory terms (for now) under realistic noise conditions (e.g. spelling variation, abbreviations, lay vs. professional language) with LLMs.
- Filter and ranking candidate synonyms for laboratory terms and construct the benchmarking dataset.
- Assess how LLM-based synonym generation affects the robustness and accuracy of terminology mapping and downstream interoperability tasks.

(Baseline 3) Normalization and pattern matching, standardizing diverse surface forms back to canonical representations through edit patterns and string transformations.

(Baseline 4) Embedding-based synonym discovery: identifying synonyms by placing terms in a vector space and retrieving nearby neighbors as candidate synonyms.

### Our approach:

Novel approach that combines (LLMs), prompt engineering, and the UMLS ontology. Diverse synonym candidates generation through prompting multiple LLMs.

## 4. Evaluation

- *Syntactic evaluation with Jaro--Winkler similarity:* we aim for this value to be low.

- *Semantic evaluation using vector embeddings and cosine similarity:* we aim for this value to be high.

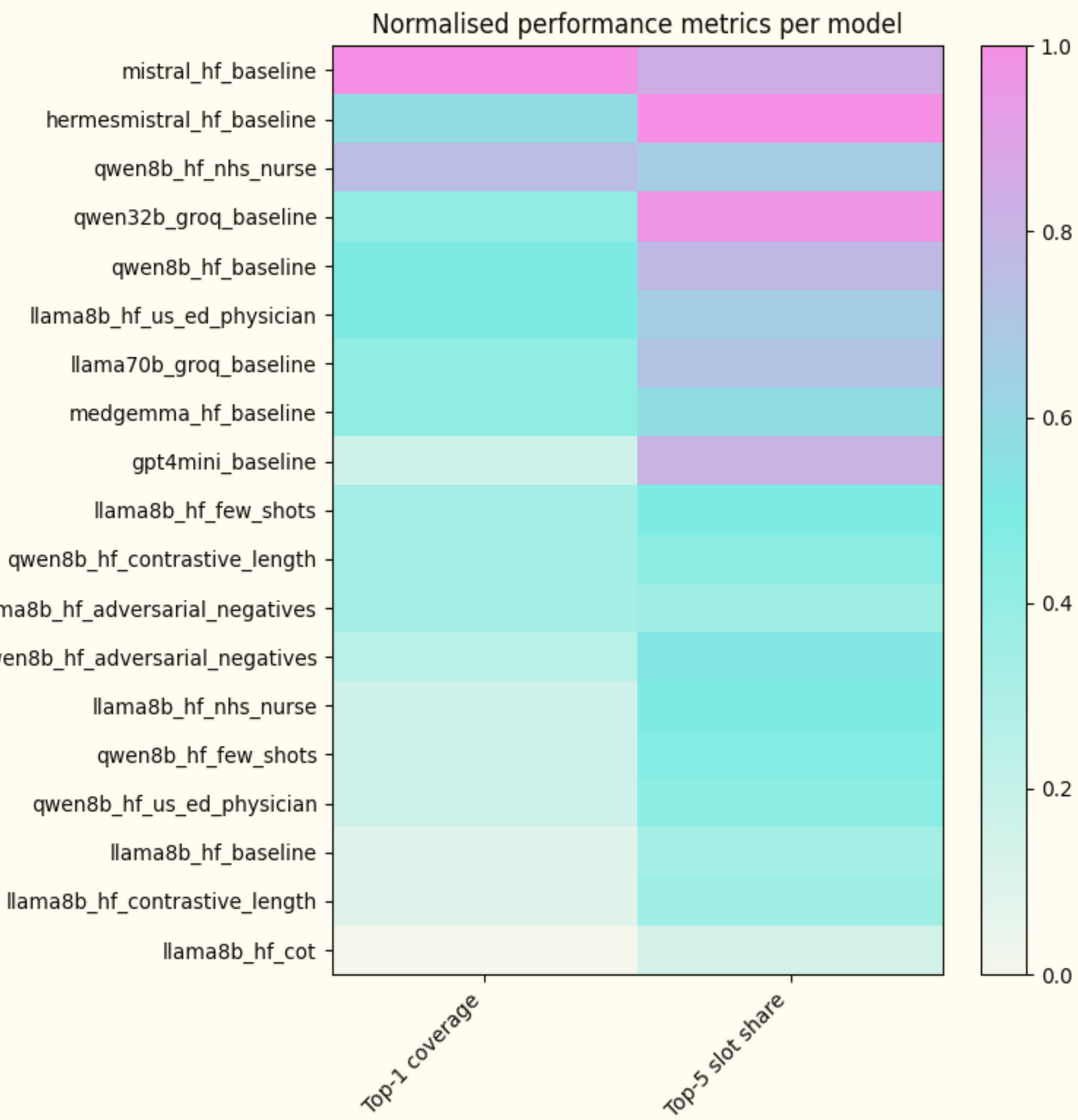
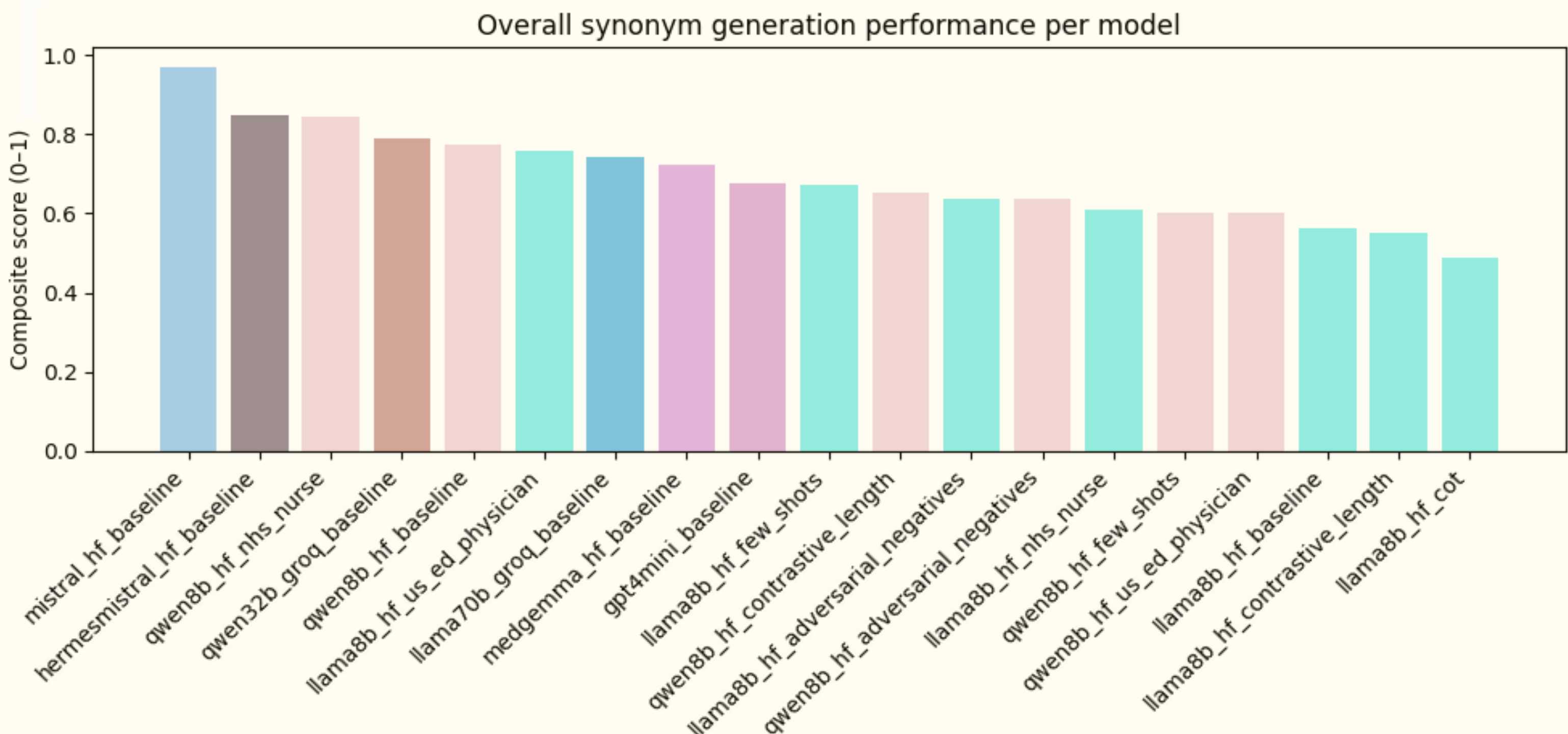
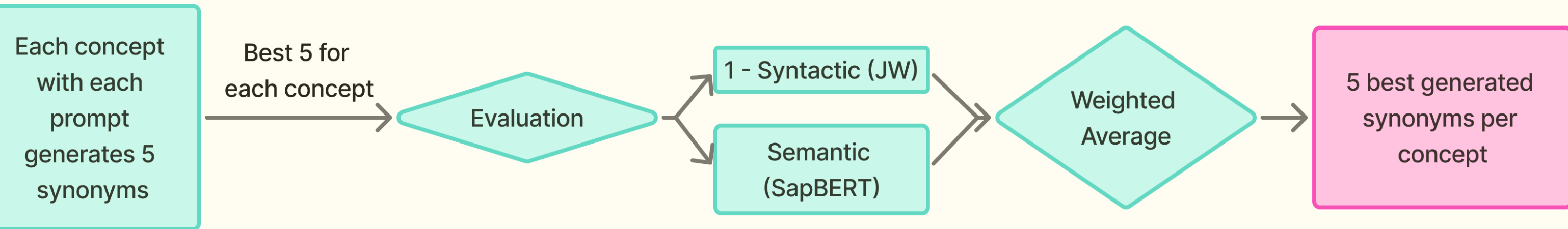
- *Clinical evaluation:* physicians will qualitatively assess the clinical correctness, plausibility and noisiness of generated synonyms (**future work**).

## 5. Preliminary Results

### Baseline vs Our Models:

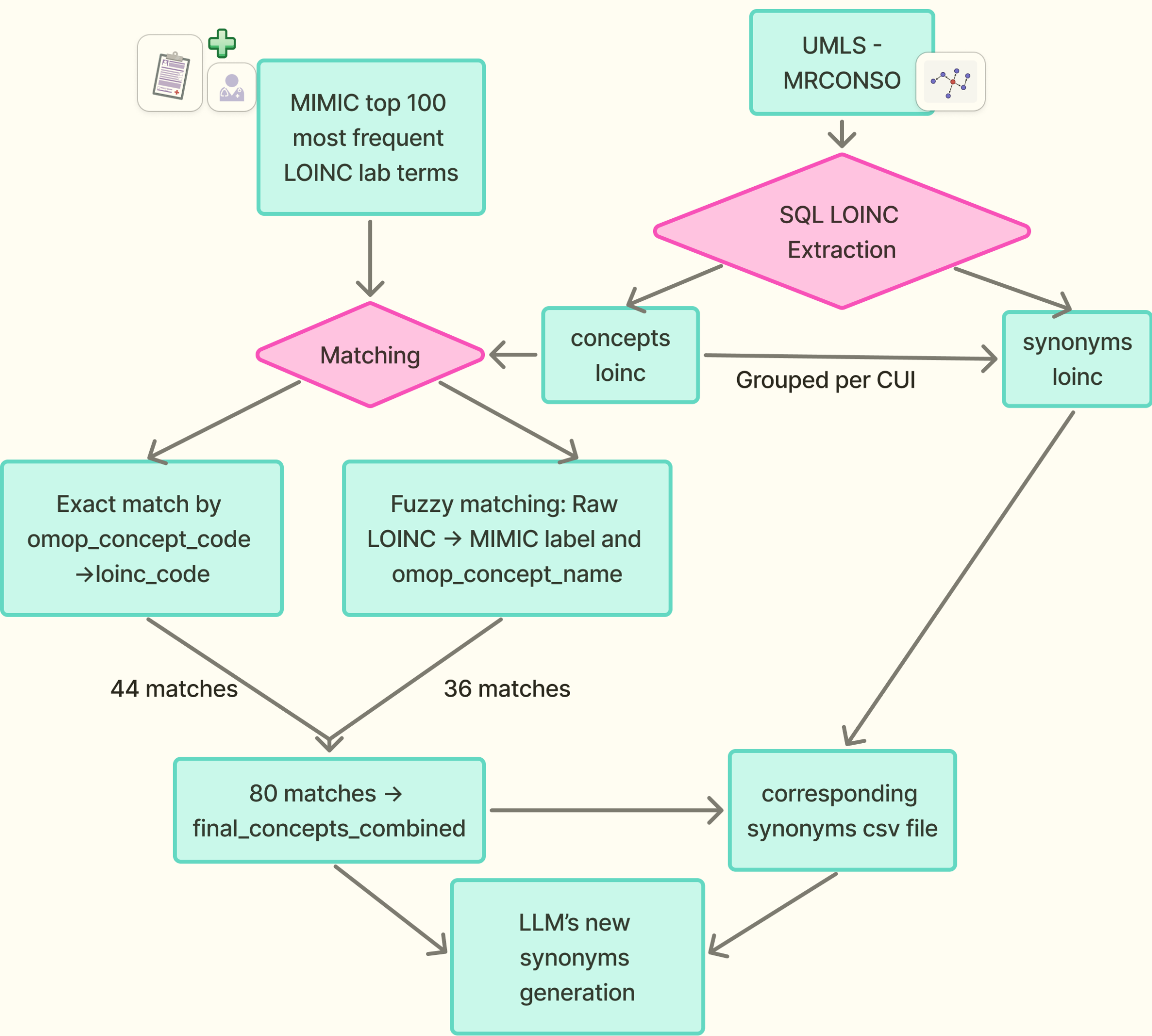
| Model                             | Mean combined score (0.6 * semantic_sim + 0.4 * (1 - syntactic_sim)) |
|-----------------------------------|----------------------------------------------------------------------|
| Baseline 1                        | 0.56                                                                 |
| Baseline 2                        | 0.51                                                                 |
| Baseline 3                        | 0.47                                                                 |
| Baseline 4                        | 0.58                                                                 |
| Our pipeline (all top-5 synonyms) | 0.63                                                                 |

### Comparison between Models:



**Examples:**  
**Bicarbonate:SCnc:Pt:Ser/Plas:Qn**  
- hco3 quant;  
- hco3 ser/pl;  
- quantitative serum bicarbonate;  
- bicarbonate serum/plasma quantitative;  
- hco3 concserpla  
**Hematocrit test status:Type:Pt:Bld:Nom:CPHS**  
- complete blood picture hematocrit status;  
- blood hematocrit test result;  
- blood cell volume test status cphs;  
- pt hct value

## 2. Input Data



## 3. Pipelines:

(Baseline 1) Lexicon and ontology-based synonym resources; curated synonyms lists (e.g. UMLS Metathesaurus, LOINC).

(Baseline 2) Rule-based variant generation: creating new synonym forms from canoni-cal ontology terms usingsyntactic rules (e.g., reordering modifiers, expanding abbreviations).

