Towards LLM-guided Healthcare Dataset Harmonization

Christos Smailis
Computational Biomedicine Lab
Department of Computer Science
University of Houston, Texas, USA
csmailis@central.uh.edu

Carlos Ordonez
Data-Intensive Parallel Algorithms for AI
Department of Computer Science
University of Houston, Texas, USA
carlos@central.uh.edu

Ioannis A. Kakadiaris
Computational Biomedicine Lab
Department of Computer Science
University of Houston, Texas, USA
ikakadia@central.uh.edu

***Abstract*—Electronic health record (EHR) datasets come in various schemas and can contain a range of data types, measurement units, and variables that share duplicate semantic content. The process of bringing such datasets into a common schema with consistent values, so that it is possible to perform queries uniformly, is known as harmonization. However, performing this process manually can be both time-consuming and prone to errors. In this work, we present a web-based platform that semi-automates the harmonization and linking of EHR datasets through a human-in-the-loop framework, guiding users with the use of large language models (LLMs). Our solution is a two-stage harmonization pipeline that keeps schema metadata processing online while handling patient-level data locally, to align with HIPAA data privacy principles. In the first stage, users harmonize and link only non-identifiable schema information. In the second stage, sensitive value-level harmonization occurs entirely on the user's system, so no private and protected health information ever leaves their environment. Throughout both stages, we expect that LLM-powered suggestions could potentially speed up the harmonization and linking processes.**

***Index Terms*—harmonization, linking, and large language models.**

I. Introduction

Linking different medical datasets enables the exchange of medical information between various medical environments, including hospitals, clinics, laboratories, and research facilities. Hospitals, research labs, and other organizations that store and handle medical datasets often employ a variety of formats (e.g., non-standardized database schemas, JSON, and CSV files) with differing naming conventions, even for similar concepts. Additionally, systems that store medical information may use different measurement units and value formats. Merging medical datasets from different institutions can offer a range of benefits, including identifying regional trends in disease prevalence and facilitating clinical and translational research through studies conducted across multiple institutions.

This work introduces a tool for harmonizing Electronic Health Record (EHR) datasets using a two-stage human-in-the-loop method accelerated by a large language model (LLM). The first stage occurs entirely through our website and allows the user to perform harmonization only for information related to dataset schemas, which describe the structure of the datasets. The second step occurs locally on the user's computer, where it harmonizes and links record values from medical datasets while preserving patient anonymity, thereby aligning with HIPAA data privacy principles.[1]

At each stage of harmonization, our system provides suggestions for establishing mappings between different table columns and for aligning value ranges across different datasets. The user can review, approve, or modify the LLM suggestions as needed.

The contributions of this work are:

- A human-in-the-loop EHR dataset harmonization pipeline that uses an LLM to assist users in aligning dataset schemas and values.
- A web-based interface that allows users to match dataset tables and attributes to perform the schema alignment process.
- A locally executable interface that allows users to match and align dataset value ranges for different attributes, ensuring no private patient information is uploaded out of the user's computer, thus aligning with HIPAA data privacy principles.

The rest of this work is organized as follows. In Section 2, we discuss related works from research literature. In Section 3, we introduce and discuss the details of our pipeline. Future experiments with our pipeline are discussed in Section 4. In Section 5, we detail the current limitations of this work, while in Section 6, we outline directions for future work. Finally, in Section 7, we discuss our conclusions.

II. Related Work

Several EHR harmonization efforts focused on converting existing medical datasets to the Observational Medical Outcomes Partnership (OMOP) Common Data Model (CDM). According to [7], OMOP CDM is a specification designed to standardize the schema and content of EHR datasets. However, not all EHR databases follow the OMOP CDM specification. Such an effort involves the conversion of the MIMIC-III v1.4 Dataset [2], which had to be performed manually, requiring up to 500 hours for two experts to complete [6], [8]. The original MIMIC-III v1.4 Dataset [3] contains EHR datasets from

[1] While our pipeline is designed to align with the HIPAA data privacy principles, compliance also depends on the user's environment.

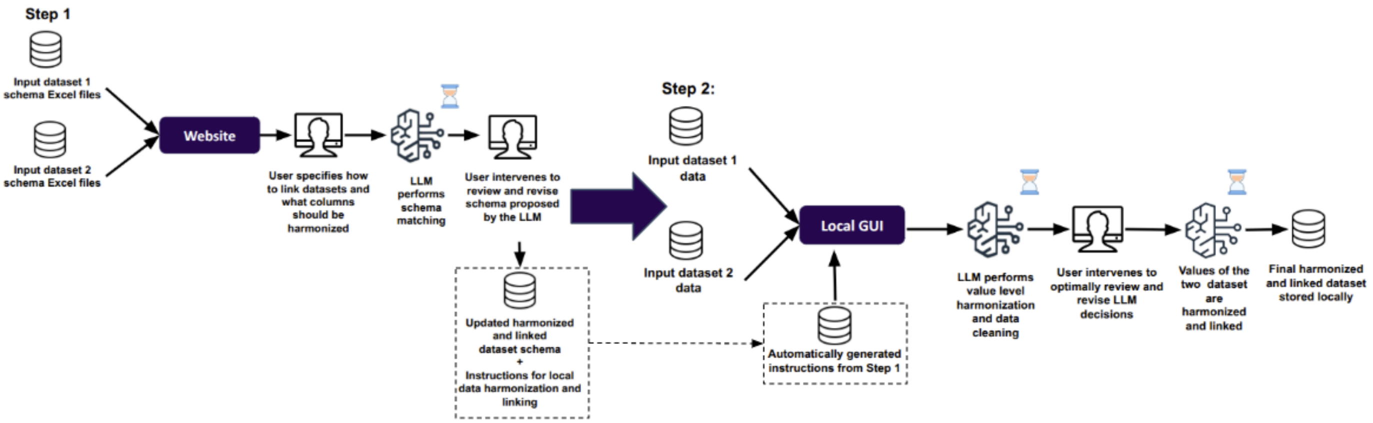
Step 1
Input dataset 1 schema Excel files
Input dataset 2 schema Excel files
Website
User specifies how to link datasets and what columns should be harmonized
LLM performs schema matching
User intervenes to review and revise schema proposed by the LLM
Updated harmonized and linked dataset schema + Instructions for local data harmonization and linking
Step 2:
Input dataset 1 data
Input dataset 2 data
Local GUI
Automatically generated instructions from Step 1
LLM performs value level harmonization and data cleaning
User intervenes to optimally review and revise LLM decisions
Values of the two dataset are harmonized and linked
Final harmonized and linked dataset stored locally

Fig. 1. Overview of our proposed harmonization pipeline: Harmonization is treated as a two-stage process that keeps schema metadata processing online while handling patient-level data locally, to align with HIPAA data privacy principles. In the first stage, users remotely harmonize and link only non-identifiable schema information. In the second stage, sensitive value-level harmonization occurs entirely on the user's own system, so no private and protected health information ever leaves their environment. Throughout both stages, LLM-powered recommendations speed up the schema and record harmonization and linking processes.

critical care units of the Beth Israel Deaconess Medical Center between 2001 and 2012. Recent methods for harmonizing EHR datasets have used either LLMs or natural language processing approaches to perform harmonization. However, they are usually limited in focusing on specific sub-problems of the EHR dataset harmonization process, either specifically processing values [4], or focusing only on the schema of different datasets [9], [10]. Many of these works also do not discuss whether or how they align with HIPAA (e.g., [9]), except [5]. In contrast, our work proposes a holistic pipeline that considers both schema and value-level harmonization, structured to align with HIPAA data privacy principles.

III. Methodology

Our harmonization pipeline features a workflow comprising two stages: In the first stage, users can harmonize the schema of two datasets. In the second stage, users can locally perform value harmonization and linking for the two datasets. As seen in Figure 1, for the first stage, a user must provide two MS Excel files that describe the schema of each dataset to be harmonized. During the first stage, no values from each dataset are uploaded to our website; instead, only anonymized schema information is uploaded, aligning with HIPAA data privacy principles. An LLM then performs schema mapping between the two datasets, and the user approves its suggestions. After the user approves the LLM's suggestions, our pipeline utilizes the LLM to generate a new schema for the harmonized dataset, along with a JSON file containing instructions on how to perform value harmonization, which can be used in stage 2. The user can download both of these files. The LLM prompt used in the current iteration of our pipeline for the first harmonization stage, which provides harmonization suggestions to the user, is shown in Fig. 2.

Fig. 2. LLM Prompt for Schema Harmonization

I need to determine if these two database columns represent the same type of information based solely on their schema metadata.

Column 1:
- Name: {col1_name}
- Table: {col1_table}
- Data Type: {col1_type}
- Description: {col1_desc}

Column 2:
- Name: {col2_name}
- Table: {col2_table}
- Data Type: {col2_type}
- Description: {col2_desc}

Based on the column names, types, and descriptions, do these columns likely contain the same type of information?

Respond with only one of these options:

- ”MATCH: [brief reason]” if they represent the same type of information
- ”NO_MATCH: [brief reason]” if they represent clearly different types of information

Consider columns as matching if they represent the same concept, even if the format differs (e.g., coded vs. text).

In the second stage of our methodology, the user must download a local client that provides a graphical user interface (GUI) for value harmonization and linking. As shown in stage 2 of Figure 1, the input to this GUI consists of the record values from the two datasets and the automatically generated value harmonization JSON file from stage 1. An assumption we make here is that the dataset records are provided in the form of CSV files, each representing a table from each of the datasets. Then an LLM checks the different unique values of the harmonized columns to establish a common mapping in case they need to be harmonized. The user gets options to perform manual value processing operations as well as those related to dataset cleaning (e.g., value replacement, date format adjustments to ensure consistency). Manual operations are performed without the assistance of the LLM and before any value harmonization mapping is established. After the LLM provides suggestions for establishing a value mapping for harmonized columns, the option to manually review and revise the suggestions is available. The LLM prompt used for value harmonization during stage 2 is shown in Fig. 3.

Fig. 3. LLM Prompt for Value Harmonization

> I have two columns with the following distinct values that need to be harmonized:
>
> Table 1 values: {table1_columns}
> Table 2 values: {table2_columns}
>
> Please suggest mappings between these values. The goal is to identify which values from Table 1 correspond to which values from Table 2.
>
> Format your response as:
> Table1: <value1>, Table2: <corresponding_value2>
>
> Only suggest mappings where you are confident the values represent the same concept. If a value has no precise match, don't include it.

After executing the above prompts, the tool parses the output returned by the LLM. It encodes the results as JSON internally to present them to the user, both during stage 1 and stage 2 of the harmonization and linking process.

The LLM is executed via Ollama either on our server (for the schema harmonization stage) for stage 1, or on the user's computer for stage 2, and does not communicate with external APIs.

IV. Future Experiments

In this section, we present an evaluation that we intend to perform in the future to assess our harmonization pipeline for aligning the schema of an EHR dataset as closely as possible with the schema of the Observational Medical Outcomes Partnership (OMOP) Common Data Model (CDM). As stated in Section 2, OMOP CDM [7] is a specification designed to standardize the schema and content of EHR datasets. However, not all EHR databases follow the OMOP CDM specification. To this end, we aim to evaluate the feasibility of using our proposed harmonization pipeline for the task of harmonizing a non-OMOP CDM-compliant EHR dataset with the specifications above.

A. Datasets

For our experiments, we intend to use the MIMIC-III Demo v1.4 Dataset [2] as the EHR dataset. This dataset comprises EHR records from the critical care units of Beth Israel Deaconess Medical Center between 2001 and 2012. The MIMIC-III Demo is a free, de-identified dataset that is a subset of the full MIMIC-III dataset [3]. However, both datasets utilize the same schema and have the same number of tables and columns. We anticipate that subsequent iterations of this work will initially focus on evaluating schema harmonization, followed by a later stage of evaluating value harmonization scenarios. Our current plan is to compare our harmonization schema coverage with the results of a previously manual attempt to harmonize the MIMIC dataset with the OMOP CDM, referred to as MIMIC-OMOP [6], [8].

B. Evaluation Metrics

To evaluate the quality of schema mapping performed by our harmonization framework, in the future, we intend to adopt the following metrics in our experiments: (i) Regarding schema harmonization, we intend to adopt metrics such as domain coverage to map the percentage of MIMIC-III Demo source tables that can be mapped to the OMOP schema. (ii) We will use column coverage to measure the amount of mapped columns from MIMIC-III against the total number of columns per table. (iii) We will adopt mapping precision to measure the percentage of correctly mapped column pairs, and mapping recall to measure the proportion of OMOP CDM table columns used in the mappings. (iv) We will use the overall schema coverage to figure out the percentage of unique OMOP CDM tables included in any mapping. (v) We intend to measure the number of user corrections that were required. Ultimately, we aim to utilize similar metrics for value-based harmonization.

We expect that our method may achieve comparable results or improved results compared to the MIMIC OMOP manual effort baseline. Similarly, we believe that it may achieve comparable or improved mapping precision, which means it may be able to discover less obvious mappings between columns and tables.

C. Implementation Details

During all harmonization stages, the user has the option to choose from a variety of LLM models provided by the Ollama platform. In the case of Stage 1, these models are executed on the server side of the website. However, in the case of stage 2, to align with HIPAA data privacy principles, the user must run the LLM models locally, as no patient records are ever transmitted during the harmonization and linking process. In our current implementation, we use the Phi-4 LLM [1] due to its state-of-the-art performance in STEM and coding tasks (taking into account its parameter number).

Our current implementation of the pipeline can be executed using either the CPU or GPU for LLM acceleration, but a GPU is preferable. We tested our implementation on a computer with an Intel(R) Core(TM) i7-10750H (12) @ 5.00 GHz CPU, 32 GB RAM, and an NVIDIA GeForce RTX 2070 GPU with 8GB of VRAM.

V. Limitations

This work, in its current form, has several limitations:

1) A critical limitation in the current iteration of this work is that the website and local GUI are currently optimized to harmonize and link two datasets at a time. This means that if a third dataset should be included in the harmonization and linking processes, it should be harmonized and linked with the outcome of the harmonization of the first two datasets. Additionally, this implies that to harmonize with OMOP CDM, a schematic description of the CDM must be provided as a CSV file.
2) The current interface of the website and local tool support performing changes to two tables of given datasets at a time.
3) Currently, our pipeline performs online schema harmonization that involves only non-sensitive user metadata. In the future, we may explore ways to perform schema harmonization collaboratively with different users. To this end, value harmonization occurs offline to align our method with HIPAA principles, ensuring that patient information remains on the user's machine. However, this may lead to situations where the LLM guiding schema harmonization may miss important information if values are not provided.
4) Finally, although we designed and developed the proposed pipeline to align with HIPAA data privacy principles, it has not yet been audited by formal compliance standards or external audits.

VI. Future Work

To address the limitations mentioned above in the future, we intend to:

1) Optimize the pipeline to directly support multi-dataset harmonization by exclusively harmonizing all prospective datasets to the OMOP CDM.
2) To improve schema harmonization, we will explore the possibility of incorporating anonymized value summaries (e.g., distributions) in this process without violating privacy.
3) It is essential to note that the LLM prompts used in the harmonization process, detailed in the methodology section above, may be revised in future iterations of this work. Additionally, future versions of this work may give users the ability to use custom prompts.

VII. Conclusions

In this work, we introduce a human-in-the-loop framework for automating EHR dataset harmonization using LLMs. Our framework enables users to perform harmonization at two stages: (i) schema harmonization, which can be performed through our online website, and (ii) value harmonization, which can be performed offline at the user's local machine, to preserve personal information that may exist in the datasets and to align with HIPAA data privacy principles. We identified limitations and highlighted future work directions. We aim to evaluate the validity of our framework through harmonizing the schema of publicly available EHR datasets with the OMOP CDM specification in future experiments.

Additionally, we intend to expand our experiments to further evaluate our method for cases involving value-level harmonization and linking scenarios.

VIII. Acknowledgments

AI text-generation tools (ChatGPT and Grammarly) were utilized in different sections of the text for editing purposes. ChatGPT and GitHub Copilot generated the text for the two LLM prompts introduced in Section 3, as we believed that LLMs can create better, more polished prompts compared to humans. The authors wrote the original manuscript, made several revisions, and edited the final draft of the document.

This work was partially funded by the National Center for Advancing Translational Sciences (NCATS), National Institutes of Health, through Grant Award Number UM1TR004539 and the UH Hugh Roy and Lillie Cranz Cullen Endowment Fund. The content is solely the responsibility of the authors and does not necessarily represent the official views of the sponsors.

References

[1] Marah Abdin, Jyoti Aneja, Harkirat Behl, Sébastien Bubeck, Ronen Eldan, Suriya Gunasekar, Michael Harrison, Russell J. Hewett, Mojan Javaheripi, Piero Kauffmann, James R. Lee, Yin Tat Lee, Yuanzhi Li, Weishung Liu, Caio C. T. Mendes, Anh Nguyen, Eric Price, Gustavo de Rosa, Olli Saarikivi, Adil Salim, Shital Shah, Xin Wang, Rachel Ward, Yue Wu, Dingli Yu, Cyril Zhang, and Yi Zhang. Phi-4 technical report. Technical report, Microsoft Research, 2024.

[2] Alistair Johnson, Tom Pollard, and Roger Mark. MIMIC-III Clinical Database Demo (version 1.4). 2019. [Online]. Available: https://doi.org/10.13026/C2HM2Q.

[3] Alistair E. W. Johnson, Tom J. Pollard, Lu Shen, Li-wei H. Lehman, Mengling Feng, Mohammad Ghassemi, Benjamin Moody, Peter Szolovits, Leo Anthony Celi, and Roger G. Mark. Mimic-III, a freely accessible critical care database. *Scientific Data*, 3:160035, 2016.

[4] Matthias Kirchler, Matteo Ferro, Veronica Lorenzini, FinnGen, Christoph Lippert, and Andrea Ganna. Large language models improve transferability of electronic health record-based predictions across countries and coding systems. medRxiv preprint, February 2025. doi: 10.1101/2025.02.03.25321597.

[5] Natallia Kokash, Lei Wang, Thomas H. Gillespie, Adam Belloum, Paola Grosso, Sara Quinney, Lang Li, and Bernard de Bono. Ontology- and LLM-based data harmonization for federated learning in healthcare. arXiv preprint arXiv:2505.20020, 2025. doi: 10.48550/arXiv.2505.20020.

[6] MIT Laboratory for Computational Physiology. MIMIC-OMOP: Mapping the MIMIC-III database to the OMOP schema. https://github.com/MIT-LCP/mimic-omop, 2018.

[7] Observational Health Data Sciences and Informatics (OHDSI). Data standardization. https://www.ohdsi.org/data-standardization/, June 2025.

[8] Nicolas Paris, Antoine Lamer, and Adrien Parrot. Transformation and evaluation of the MIMIC database in the OMOP Common Data Model: Development and usability study. *JMIR Medical Informatics*, 9(12):e30970, 2021.

[9] Nabeel Seedat and Mihaela van der Schaar. Matchmaker: Self-improving large language model programs for schema matching. arXiv preprint arXiv:2410.24105, 2024. doi: 10.48550/arXiv.2410.24105.

[10] Xinyu Zhou, Lovedeep Singh Dhingra, Arya Aminorroaya, Philip Adejumo, and Rohan Khera. A novel sentence transformer-based nlp approach for schema mapping of electronic health records to the OMOP common Data Model. pages 1332–1339, Atlanta, GA, USA, May 2025.