Research Proposal: Using LLM Agents for Medical Data Cleaning

Siqi Yao

Contents

1 Introduction

Data cleaning is a crucial yet tedious task, which has led to the pursuit of autonomous solutions, with LLM agents emerging as a promising approach. For instance, [BDH25] provides an in-depth exploration of how LLM agents can be used to automate the entire data cleaning process, and evaluates the proposed method across three Kaggle datasets—Titanic, Meat Consumptions, and Hotel Bookings—to assess whether data cleaning through agents alone can enhance the performance of machine learning models. [LFLT24] purposes a LLM-based pipeline for automatically generating data-cleaning workflows (OpenRefine operations), and [QMW24] automates data standardization process by proposing a Python library with declarative, unified APIs for standardization, and combining it with LLM agents.

Many public medical / healthcare-related datasets are dirty but are widely used in many papers. Thus, cleaning medical datasets is of paramount importance. [DVN+22] discover that many previously published dermatological AI algorithms rely on images labeled by visual consensus, which can be noisy since the dermatologists who labeled them lack necessary information used for diagnosis. This may mislead diagnosis, even resulting in increased cost or mortality. Thus, the paper created the Diverse Dermatology Images (DDI) dataset—the first publicly available, expertly curated, and pathologically confirmed image dataset with diverse skin tones.

Motivated by this idea, I searched for papers specifically focusing on the use of agents for cleaning medical data, but could not find any. I only discovered three relevant papers. [WZW+24] and [SVG+25] both discuss the construction of agents for end-to-end analysis of healthcare data, with data cleaning being just one step in the process. Meanwhile, [SPLF25] explores how LLM agents can be utilized for data harmonization, which involves merging data from different sources, using healthcare data as an example. Hence, my proposed research topic is **Using LLM Agents for Medical Data Cleaning**.

2 Methodology

The core of our proposed methodology is the design, implementation and evaluation of a specialized LLM agentic system tailored for the unique challenges of medical data cleaning. The overall architecture of our system is illustrated in Figure 1.

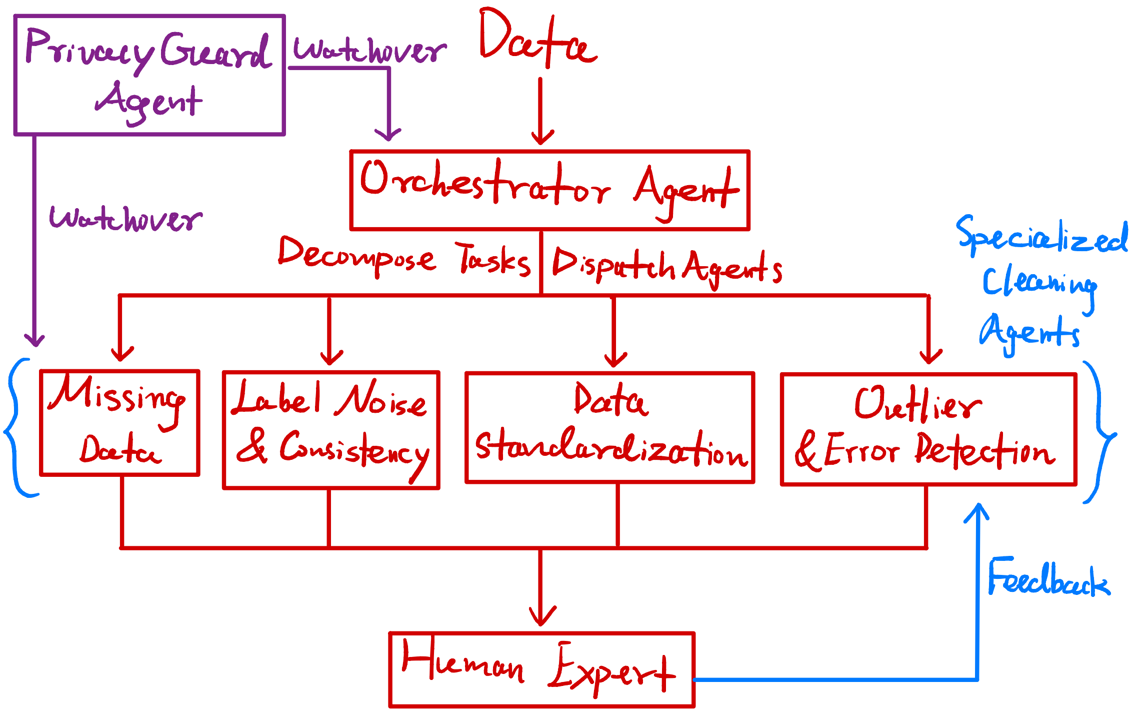

Figure 1: Complete Architecture.

2.1 Agent System Design and Development

We will design a multi-agent system where each agent or agent group is specialized for a specific data cleaning task and equipped to handle medical data constraints.

1. **Agent Orchestrator**: A central controller agent will be responsible for the overall workflow. Its tasks include:

    - **Initial Data Profiling**: Perform an initial analysis of the dataset to understand its schema, data types and a preliminary assessment of data quality issues (e.g., missing value rates, class distribution, unique value analysis).
    - **Task Decomposition**: Based on the profile, decompose the cleaning process into subtasks (e.g., " handle missing values in column ' blood_pressure ' ").
    - **Agent Dispatch**: Assigning these subtasks to the most appropriate specialized agent.
    - **Context Management**: Maintaining a shared context to ensure decisions are consistent across the dataset (e.g., if one agent standardizes " Heart Disease " to " HD ", all other agents should be aware).

2. **Specialized Cleaning Agents**: We will develop a group of agents, each with a specific goal:

    - **Missing Data Agent**: tasked with handling missing values. Instead of simple imputation, it will use LLM reasoning to suggest context-aware solutions. For example, it can distinguish between Missing Completely at Random (MCAR) and Missing Not at Random (MNAR) (e.g., a missing " tumor_size " might be because it was not measurable, which is itself a critical clinical insight). It will propose actions like creating a new indicator variable for " missingness " or using sophisticated imputation methods guided by medical knowledge.

- **Label Noise & Consistency Agent**: focused on categorical variables and class labels. It will leverage the LLM's embedded medical knowledge to identify implausible or inconsistent labels (e.g., a diagnosis of " Type 1 Diabetes " for a 70-year-old patient might be marked for review). For image data, it could generate prompts for a vision-language model to verify if an image's content matches its label (e.g., " Does this X-ray image show a bone fracture? ").
- **Outlier & Error Detection Agent**: tasked with identifying numerical and textual outliers. It will go beyond statistical metrics by using semantic checks. For instance, a blood pressure reading of " 1850/110 " or a patient's age of " 200 " is clearly invalid. The agent will be prompted to identify such errors based on clinical plausibility.
- **Data Standardization Agent**: responsible for harmonizing formats (e.g., standardizing date formats to `DD/MM/YYYY`), units (e.g., converting pounds to kg for weight), and medical terminologies (e.g., mapping " Heart Disease " to " HD ").

3. **Privacy and Compliance Guard**: A critical component unique to medical data.

    - This module will intercept all data before it is sent to an external LLM API (e.g., GPT-4, Claude).
    - It will perform automatic de-identification, stripping privacy information such as names, birth dates and phone numbers.
    - For operations requiring external APIs, it will employ techniques like synthetic data generation or federated learning setups to ensure raw data never leaves the secure environment.
    - It will enforce a strict data governance policy, logging all accesses and transformations.

2.2 Experimental Setup and Evaluation

1. **Data Selection**: To show the generalizability of our method, besides **real-world use cases from hospitals**, we will include nine datasets drawn from three different domains, with three datasets from each domain. The datasets should be publicly available, widely-used, but known-to-be-noisy (for most choices). Potential candidates include:

    - Dermatology
        - **The Diverse Dermatology Images (DDI) dataset** [DVN+22].
        - **The DermNet dataset** [GH20]: Contains 19,500 images representing 23 different types of skin diseases, with a number of irrelevant images included.
        - **The Fitzpatrick17k dataset** [GHS+21]: Contains 16,577 clinical images with skin condition labels and skin type labels.
    - Chest X-ray
        - **The NIH Chest X-rays dataset** [Cra17]: Comprised of 112,120 X-ray images with disease labels from 30,805 unique patients. The image labels are NLP extracted so there could be some erroneous labels but the NLP labeling accuracy is estimated to be $> 90\%$.
        - **The CheXpert Dataset** [IRK+19]: Contains 224,316 chest radiographs of 65,240 patients labeled for the presence of 14 common chest radiographic observations. Designed labels that represent uncertainty, which can be viewed as a kind of noise.
        - **The MIMIC-CXR and MIMIC-CXR-JPG Dataset** [JPB+19, JPG+19]: MIMIC-CXR contains 227,835 imaging studies for 65,379 patients presenting to the Beth Israel Deaconess Medical Center Emergency Department between 2011–2016.

MIMIC-CXR-JPG contains 377,110 chest x-rays associated with 227,827 imaging studies sourced from the same location and time span.

- EHR / ICU
    - **The MIMIC-III Dataset** [JPS+16]: a large, single-center database comprising information relating to patients admitted to critical care units at a large tertiary care hospital.
    - **The eICU Collaborative Research Database** [PJR+18]: A multi-center ICU database with high granularity data for over 200,000 admissions to ICUs monitored by eICU Programs across the United States. The database is deidentified, and includes vital sign measurements, care plan documentation, severity of illness measures, diagnosis information, treatment information, and more.
    - **The HiRID Dataset** [FZL+21]: A freely accessible critical care dataset containing data relating to more than 33 thousand admissions to the Department of Intensive Care Medicine of the Bern University Hospital, Switzerland, an interdisciplinary 60-bed unit admitting ¿6,500 patients per year.

2. **Baseline Methods**: The performance of our LLM Agentic system will be compared against:
    - **Traditional Rule-based Methods**: Scripts based on predefined rules and thresholds.
    - **Standard Automated Tools**: Such as Python libraries like pandas or commercial data cleaning tools.
    - **Ablation Study**: We will run versions of our system with certain agents disabled to isolate the contribution of each component (e.g., system performance without the Label Noise Agent).

3. **Evaluation Metrics**: Success will be measured by two sets of metrics:
    - **Data Quality Metrics**: This will include metrics such as completeness, which measures the reduction in missing values; consistency, which tracks the rate of resolved inconsistencies; and plausibility, which involves expert evaluation of the clinical plausibility of imputed values or corrected labels.
    - **Downstream Utility Metrics**: The ultimate test of data cleaning is its impact on analysis. We will train machine learning models on the raw, baseline-cleaned and agent-cleaned datasets and compare model performance such as F1-Score and AUC-ROC.

2.3 Expert Assistance and Validation

Realizing that full automation in medicine is risky, we plan to incorporate a human validation step.

- The system will generate a report for each change it proposes, including:
    - The original value.
    - The proposed change/correction.
    - The reasoning from the agent.
- This report will be presented to a domain expert for approval, rejection or modification. This feedback can also be used to fine-tune the agent's decision-making process iteratively.

We will also give careful consideration to the following ethical aspects:

- **Privacy**: This is the foremost concern. We will try to use local open-source LLMs (e.g., Llama 3) to avoid sending data to third parties. When using APIs, we will strictly use their privacy-preserving options and never send private information.

- **Bias**: We are aware that LLMs can inherit and amplify biases present in their training data. We will carefully prompt the agents to be fair and will evaluate the cleaned data for potential introduced biases across different demographic groups.
- **Hallucination**: LLMs can sometimes hallucinate or generate incorrect but plausible-sounding information. The expert-in-the-loop validation is crucial safeguard against this.

3 Expected Outcomes

In summary, this research plans to develop a specialized LLM agentic framework for cleaning medical datasets, integrating error-specific agents, privacy protection and human validation. We aim to show its superiority over traditional methods, improving data quality and machine learning model performance. Additionally, we will uncover previously overlooked issues of public medical datasets, challenging current benchmarks and encouraging more rigorous data cleaning for reliable medical AI research.

References

[BDH25] T. Bendinelli, A. Dox, and C. Holz. Exploring llm agents for cleaning tabular machine learning datasets. *arXiv preprint arXiv:2503.06664*, 2025.

[Cra17] C. Crawford. Nih chest x-rays. https://www.kaggle.com/datasets/nih-chest-xrays/data/data, 2017.

[DVN+22] R. Daneshjou, K. Vodrahalli, R. A. Novoa, M. Jenkins, W. Liang, V. Rotemberg, J. Ko, S. M. Swetter, E. E. Bailey, O. Gevaert, et al. Disparities in dermatology ai performance on a diverse, curated clinical image set. *Science advances*, 8(31):eabq6147, 2022.

[FZL+21] M. Faltys, M. Zimmermann, X. Lyu, M. Hüser, S. Hyland, G. Rätsch, and T. Merz. Hirid, a high time-resolution icu dataset (version 1.1. 1). *Physio. Net*, 10, 2021.

[GH20] S. Goel and B. Hall. Dermnet. https://www.kaggle.com/datasets/shubhamgoel27/dermnet, 2020.

[GHS+21] M. Groh, C. Harris, L. Soenksen, F. Lau, R. Han, A. Kim, A. Koochek, and O. Badri. Evaluating deep neural networks trained on clinical images in dermatology with the fitzpatrick 17k dataset. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1820–1828, 2021.

[IRK+19] J. Irvin, P. Rajpurkar, M. Ko, Y. Yu, S. Ciurea-Ilcus, C. Chute, H. Marklund, B. Haghgoo, R. Ball, K. Shpanskaya, et al. Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pages 590–597, 2019.

[JPB+19] A. E. Johnson, T. J. Pollard, S. J. Berkowitz, N. R. Greenbaum, M. P. Lungren, C.-y. Deng, R. G. Mark, and S. Horng. Mimic-cxr, a de-identified publicly available database of chest radiographs with free-text reports. *Scientific data*, 6(1):317, 2019.

[JPG+19] A. E. Johnson, T. J. Pollard, N. R. Greenbaum, M. P. Lungren, C.-y. Deng, Y. Peng, Z. Lu, R. G. Mark, S. J. Berkowitz, and S. Horng. Mimic-cxr-jpg, a large publicly available database of labeled chest radiographs. *arXiv preprint arXiv:1901.07042*, 2019.

[JPS$^+$16] A. E. Johnson, T. J. Pollard, L. Shen, L.-w. H. Lehman, M. Feng, M. Ghassemi, B. Moody, P. Szolovits, L. Anthony Celi, and R. G. Mark. Mimic-iii, a freely accessible critical care database. *Scientific data*, 3(1):1–9, 2016.

[LFLT24] L. Li, L. Fang, B. Ludäscher, and V. I. Torvik. Autodcworkflow: Llm-based data cleaning workflow auto-generation and benchmark. *arXiv preprint arXiv:2412.06724*, 2024.

[PJR$^+$18] T. J. Pollard, A. E. Johnson, J. D. Raffa, L. A. Celi, R. G. Mark, and O. Badawi. The eicu collaborative research database, a freely available multi-center database for critical care research. *Scientific data*, 5(1):1–13, 2018.

[QMW24] D. Qi, Z. Miao, and J. Wang. Cleanagent: Automating data standardization with llm-based agents. *arXiv preprint arXiv:2403.08291*, 2024.

[SPLF25] A. Santos, E. H. Pena, R. Lopez, and J. Freire. Interactive data harmonization with llm agents: Opportunities and challenges. *arXiv preprint arXiv:2502.07132*, 2025.

[SVG$^+$25] S. R. Shimgekar, S. Vassef, A. Goyal, N. Kumar, and K. Saha. Agentic ai framework for end-to-end medical data inference. *arXiv preprint arXiv:2507.18115*, 2025.

[WZW$^+$24] H. Wu, Y. Zhu, Z. Wang, X. Zheng, L. Wang, W. Tang, Y. Wang, C. Pan, E. M. Harrison, J. Gao, et al. Ehrflow: A large language model-driven iterative multi-agent electronic health record data analysis workflow. In *KDD'24 Workshop: Artificial Intelligence and Data Science for Healthcare: Bridging Data-Centric AI and People-Centric Healthcare*, 2024.