Computers in Biology and Medicine 192 (2025) 110181

Contents lists available at ScienceDirect

Computers in Biology and Medicine

journal homepage: www.elsevier.com/locate/compbiomed

Zero-shot learning for clinical phenotyping: Comparing LLMs and rule-based methods

Bernardo Neves [a,b,c,d,*], José Maria Moreira [a], Simão Gonçalves [a], Jorge Cerejo [a], Nuno A. da Silva [a], Francisca Leite [a,d], Mário J. Silva [c]

[a] Hospital da Luz Learning Health, Luz Saúde, Lisboa, Portugal
[b] Internal Medicine Department, Hospital da Luz Lisboa, Lisboa, Portugal
[c] INESC-ID, Instituto Superior Técnico, Universidade de Lisboa, Portugal
[d] Católica Medical School, Universidade Católica Portuguesa, Portugal

ARTICLE INFO

Keywords:
Phenotyping
Multimorbidity
Zero-shot learning
Large language models

ABSTRACT

Background: Phenotyping, the process of systematically identifying and classifying conditions within clinical data, is a crucial first step in any data science work involving Electronic Health Records (EHRs). Traditional approaches require extensive manual annotation efforts and face challenges with scalability.
Methods: We investigated the use of Large Language Models (LLMs) for zero-shot phenotyping of 20 prevalent chronic conditions based on synthetic patient summaries generated from real structured EHRs codes. We evaluated the performance of multiple LLMs, including GPT-4o, GPT-3.5, and LLaMA 3 models with 8-billion, 70-billion, and 405-billion parameters, comparing them against traditional rule-based methods. For the analysis we used a dataset of 1,000 patients from Hospital da Luz Lisboa.
Results: GPT-4o outperformed both traditional rule-based methods and alternative LLMs, achieving superior recall (0.97) and macro-F1 score (0.92). Rule-based phenotyping, while highly precise (0.92), showed lower recall (0.36). The integration of rule-based methods with LLMs optimized phenotyping accuracy by targeting manual annotation efforts on discordant cases.
Conclusion: Zero-shot learning with LLMs, particularly GPT-4o, offers a powerful and efficient approach for phenotyping chronic conditions from EHRs, significantly reducing the need for extensive labeled datasets while maintaining high accuracy and interpretability.

1. Introduction

Multimorbidity, defined as the coexistence of multiple conditions within the same individual, is increasingly recognized as a critical public health issue [1]. The term implies the coexistence of at least two chronic conditions, though the aim is often to capture more complex patients with high needs [2,3]. Although there is no consensus on a list of chronic conditions to be included in the definition, studies commonly suggest analyzing dozens of broad-level categories of conditions [4,5]. Electronic Health Records (EHRs) provide a comprehensive, longitudinal view of patient health, capturing detailed information on multiple conditions, treatments, and outcomes, thus serving as a valuable data source for epidemiological and clinical research of multimorbidity. However, their analysis poses several challenges, such as the incompleteness of records, differing terminologies for concepts, heterogeneity of data standards, and abundant unstructured data, such as clinical text, which presents additional challenges for analysis [6].

Phenotyping, the process of identifying patients with observable traits or characteristics of interest, involves more than just assigning a list of codes to a patient record. It requires extracting clinically relevant features from patient data to create cohorts with specific characteristics [7,8]. Rule-based approaches, which rely on iterative queries and clinical expertise to map variables of interest, are highly interpretable but time-consuming to code, hard to maintain, and prone to biases [7,9]. Traditional machine learning methods, particularly supervised learning, reduce expert effort but require high-quality labeled data, which is often difficult to obtain [10]. Deep learning techniques, which use representation learning to reduce the need for extensive supervision, have advanced data-driven phenotyping and disease subtyping. However, their clinical implementation faces challenges due to difficulties in assessing algorithm accuracy and ensuring interpretability [11–13].

* Corresponding author at: Hospital da Luz Learning Health, Luz Saúde, Lisboa, Portugal.
E-mail address: bernardoneves@tecnico.ulisboa.pt (B. Neves).

https://doi.org/10.1016/j.compbiomed.2025.110181
Received 14 September 2024; Received in revised form 14 March 2025; Accepted 7 April 2025
Available online 23 April 2025

In recent years, Transformer-based machine learning models, have become the state-of-the-art in natural language processing (NLP) [14]. Among them, large language models (LLMs), which are complex deep learning models trained on vast text corpora, have achieved remarkable performance across benchmarks in several different NLP tasks [15]. In the medical domain, LLM-based chatbots have already shown remarkable capabilities in tasks like clinical text summarization and answering questions on medical licensing exams [16,17]. A few studies have explored the use of general domain-trained or medically domain-trained language models for phenotype extraction, primarily utilizing encoder-based models like BERT [18]. While this approach has improved over previous methods, it still requires significant data labeling and fine-tuning, which can become problematic when multiple phenotypes are needed or in the presence of rare phenotypes.

A few recent works have begun to explore the use of larger decoder-based models for phenotype extraction from EHRs without the need for extensive fine-tuning on annotated data. Alenszer et al. showed the usefulness of Flan-T5 model to perform zero-shot extraction of obstetric phenotypes-[19]. Wei et al. introduced a two-step framework in which GPT-4 drafts phenotyping SQL queries of EHR data organized under the Observational Medical Outcomes Partnership (OMOP) Common Data Model (CDM) [20]. Subsequent expert validation is still required to resolve logical inconsistencies in inclusion and exclusion criteria, but this hybrid approach reduced development time by up to 40%. Further studies demonstrated that through more advanced prompt engineering, such as chain-of-thought reasoning, hallucinations could be reduced and phenotyping performance improved even in scenarios where training examples are rare or absent [21]. Despite these advancements, research has largely focused on a limited range of phenotyping cases without systematic evaluation across conditions [22,23], while the integration of LLMs into traditional workflows to optimize annotation and large-scale extraction remains underexplored.

In this work, we propose using zero-shot learning for phenotyping EHR data, leveraging state-of-the-art LLMs to classify a wide range of chronic conditions without extensive labeled datasets. This approach reduces manual annotation compared to traditional rule-based methods, which, although sensitive, often lack specificity. Our contributions include: (1) the first comprehensive evaluation of zero-shot LLM-based clinical phenotyping across multiple chronic conditions; (2) an innovative hybrid method that combines rule-based techniques with LLM predictions to optimize annotation efforts; (3) detailed analyses of model confidence and rationale generation, offering new insights into LLM interpretability in clinical applications; and (4) a novel evaluation of fairness and bias across demographic subgroups. Unlike previous studies that focused on specific conditions or required extensive fine-tuning, our method demonstrates robust performance without condition-specific training data.

2. Methods

This study involved a multi-step process, as shown in the flowchart in Fig. 1. It included data extraction from EHRs converted into the OMOP CDM, data sampling, synthetic patient summary generation, rule-based dictionary creation, prompt development, phenotyping evaluation, annotation of discordant cases and assessment of classification metrics.

2.1. Data preprocessing

The data used in this study was sourced from Hospital da Luz Lisboa, the largest private hospital in Portugal and a major medical institution in Lisbon [24]. The dataset includes anonymized patient data spanning from 2007 to 2023 that was extracted from various hospital systems and then transformed to conform to the OMOP CDM version 5.4. The OMOP CDM data was produced within IntelligentCare, a project that aimed to develop a patient-centric solution for managing multimorbidity using analytical methods to explore EHR data and patient-reported measures, leveraging smart sensors and mobile solutions [25].

We generated synthetic patient summaries from the text descriptions of OMOP codes for conditions, procedures, laboratory measurements, and drug tables, along with demographic variables in EHRs (see Fig. 1). We retrieved the text descriptions associated to every OMOP code to obtain clinical texts in English and created structured JSON files that are both human-readable and facilitate further processing (an example is provided in Fig. 2). The data is organized by pseudonymized patient ID and sex, with events listed in the chronological order, based on the patient's age at the time. Each event includes details such as event ID, date, conditions, procedures, laboratory measurements, and drug prescriptions. This structured format allows for a comprehensive view of each patient's medical history, making it suitable for both manual review and machine learning applications (see Fig. 3).

The data preprocessing and transformation were performed using *Pyspark* in *Azure Databricks*. The prompting and evaluation was conducted using *Python 3.8*.

2.2. Chronic conditions analyzed and rule-based dictionary creation

A group of six physicians from different specialties (Internal Medicine, Intensive Care, Cardiology and Clinical Pharmacology), each with more than 10 years of clinical expertise, initially selected a list of chronic conditions through a consensus. The final list of 55 conditions was consistent with other reported lists in the literature [4,26]. Due to feasibility and budget constraints, we focused on the 20 most prevalent conditions on the overall dataset. For each of these conditions, we built a comprehensive dictionary encompassing conditions, procedures, laboratory measurements, and drug prescriptions. We used Observational Health Data Science and Informatics (OHDSI) standardized vocabularies and comprehensive terminologies to define condition and procedure codes (standard OMOP concept ids) for each chronic condition, ensuring that all descendant codes were included in the definition. For drug prescriptions, we adapted a previously published mapping from Anatomical Therapeutical Chemical classification system (ATC) codes to conditions [27]. Additionally, heart failure, diabetes mellitus, hyperlipidemia, and anemia were mapped from laboratory measurements according to standard definitions. The final dictionary is included in Appendix.

2.3. Models assessed and prompting strategies

2.3.1. Data selection

Given the highly sparse nature of EHRs, where only a small fraction of patient records mention chronic conditions, we first filtered the dataset to include only patients flagged as having at least one of the 55 conditions defined in our dictionary-based rules. This step ensured that our analysis focused on a clinically relevant subset of patients. From this pre-filtered cohort, we then randomly sampled 1,000 adult patients for further analysis. These 1,000 patients collectively had 17,543 clinical events, each of which was evaluated for the presence of a specific subset of 20 chronic conditions, resulting in 20,000 separate evaluations. This sampling strategy balanced the need for a representative dataset with the feasibility of manual annotation and the computational cost of LLM evaluations.

2.3.2. Prompt development

Zero-shot learning is a paradigm where models generalize to new tasks without additional task-specific training, relying solely on their pre-existing knowledge [28]. Leveraging this capability through prompt learning, we developed a structured prompt to classify chronic conditions based on clinical notes. Our approach draws significant inspiration from the HealthPrompt framework—a methodology specifically tailored for clinical NLP tasks [29]. HealthPrompt emphasizes the

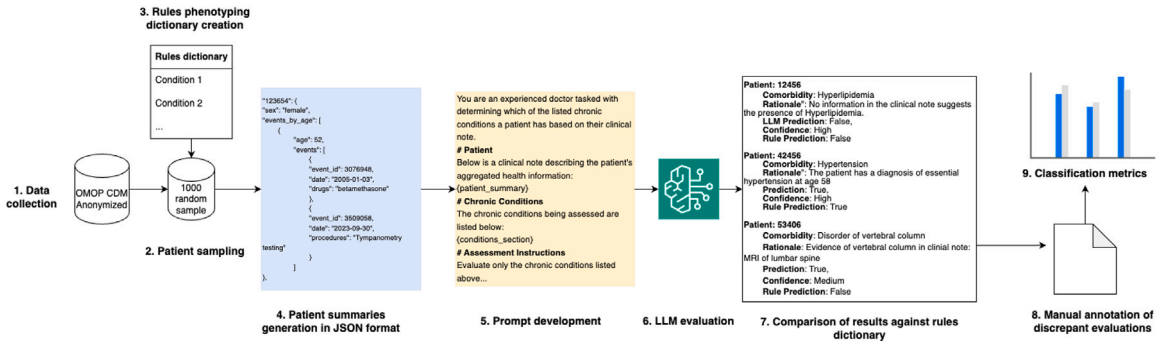


Fig. 1. Workflow of the study methodology. The steps are as follows: (1) Data Collection from Hospital da Luz Lisboa, spanning from 2007 to 2023. (2) Random sampling of 1000 patients. (3) Rule-based Dictionary Creation for chronic conditions using standardized vocabularies. (4) Synthetic patient summary generation from text conversion of OMOP codes descriptions in English and creation of a JSON file. (5) Prompt development. (6) Model Evaluation using GPT-3.5 and GPT-4 via secure Azure PHI-compliant instances. (7) Comparison of LLM classifications against Rule-based Dictionary to identify discrepancies. (8) Manual Annotation of discrepancies by a physician. (9) Performance Metrics Calculation using precision, recall, and F1 score.


```
{
    "patient_id": "296888",
    "sex": "female",
    "events_by_age": [
        {
            "age": 41,
            "events": [
                {
                    "event_id": 3418651,
                    "date": "2009-09-30",
                    "conditions": "Injury of free lower limb",
                    "procedures": "Skeletal X-ray of lower limb, Radiography of ankle"
                    "measurements": "Bilirubin.total: 0.74, Alkaline phosphatase: 103.0, MCHC: 28.0,
                         Chloride: 105.0, Gamma glutamyl transferase: 23.0, Hemoglobin: 13.5, Cancer
                         Ag 125: 7.2, Leukocytes: 5.9, Erythrocyte distribution width: 15.3, Urea: 50
                         .0, Glucose: 115.0, Sodium: 141.0, Creatinine: 0.64, Potassium: 5.2,
                         Hematocrit: 41.3, Aspartate aminotransferase: 16.0, Platelets: 242.0, MCH: 3
                         3.0, Alanine aminotransferase: 17.0, Erythrocytes: 4.8, Lactate
                         dehydrogenase: 190.0, MCV: 86.0, Bilirubin.conjugated: 0.22"
                },
        {
            "age": 42,
            "events": [
                {
                    "event_id": 9378567,
                    "date": "2010-07-01",
                    "procedures": "Arthrotomy, Arthroscopy",
                    "drugs": "diclofenac, thiocolchicoside"

                }
            ]
        }
    ]
}
```


Fig. 2. Example of a hypothetical patient summary organized as a JSON data structure. Conditions, procedures, laboratory measurements, and drug prescription codes are converted into their text descriptions and arranged chronologically.

use of structured templates that harness zero-shot learning to extract nuanced clinical insights from free-text data. By integrating clear, domain-specific instructions and specifying a structured output format (a JSON response), our design minimizes ambiguity and enhances the reliability of generated outputs, ensuring that all relevant clinical information is considered when evaluating chronic conditions.

2.3.3. LLMs selection and configuration

For our assessment, we utilized two proprietary models from OpenAI: GPT-3.5 (model version 0613) [30] and GPT-4o (model version 2024-05-13) [31], as well as three open-source models from Meta: LLaMA 3.1 8b, 70b, and 405b Instruct [32]. The choice of these models was driven by their ready availability through the Azure cloud provider, which our institution authorizes for data use. All models were configured with a temperature of 0.5 and a maximum token limit of 5,000 to generate responses, balancing creativity with coherence and ensuring detailed output within a specified length.

2.4. Annotation process and phenotyping strategy evaluation

After the preprocessing steps above described, the concatenation of the prompt and each patient summary was input to the LLMs to determine the presence of each of the 20 conditions. The models generated a true or false answer regarding the presence of each condition, accompanied by an explanation. Explanations included whether the condition was observed, the rationale behind the assessment, and the confidence level (low, medium, high). All patients were classified both by the LLMs and the rules dictionary.

We compared the classifications of all conditions in each patient between all the LLMs dictionary rules, flagging cases with discrepancies. A physician manually reviewed all cases where there was an observed discrepancy and assigned a ground truth classification of such cases. For the remaining assessments where both rules validation and LLMs agreed, the ground truth was automatically established. We randomly selected 100 concordant cases for human annotation to verify agreement. The annotation process is outlined in Fig. 4. In addition, we separately classified the presence of condition codes alone—specifically, the condition codes used in the dictionary rules—while excluding the

```
# Task
You are an experienced doctor tasked with determining which of the listed chronic conditions a patient has based on their clinical note.

# Patient
Below is a clinical note describing the patient's aggregated health information:
{patient_summary}

# Chronic Conditions
The chronic conditions being assessed are listed below:
{conditions_section}

# Assessment Instructions
Evaluate only the chronic conditions listed above. Do not consider or mention any other conditions.

For each chronic condition, use the patient's clinical note to determine whether the patient has that chronic condition.
Provide a detailed explanation.

First, consider any related medical terms, subtypes, or diagnoses related to the chronic condition. If any related terms or subtypes are found,
confirm the presence of the condition.
Then, look at the clinical note sections such as procedures, measurements, and medications to infer if there could be a chronic condition based
on the procedures, laboratory results, and medications.
Remember to consider subtypes of the chronic conditions when making your assessment.

Provide your response as a JSON list of dictionaries, where each dictionary contains the following elements:
* comorbidity: str - The name of the comorbidity being assessed
* rationale: str - Your reasoning for the assessment
* is_met: bool - "true" if the patient has the comorbidity, otherwise "false"
* confidence: str - Your confidence level ("low", "medium", "high")

An example of how your JSON response should be formatted is shown below, where the list of JSON dictionaries is stored in the "assessments" key:
'''json
{
"assessments" : [
    {
    "comorbidity" : "comorbidity_1",
    "rationale" : "Reason for assessment",
    "is_met" : true/false,
    "confidence" : "low/medium/high"
    },
    {
    "comorbidity" : "comorbidity_2",
    "rationale" : "Reason for assessment",
    "is_met" : true/false,
    "confidence" : "low/medium/high"
    }
]
}
```

Fig. 3. Prompt for determining chronic conditions based on clinical notes.

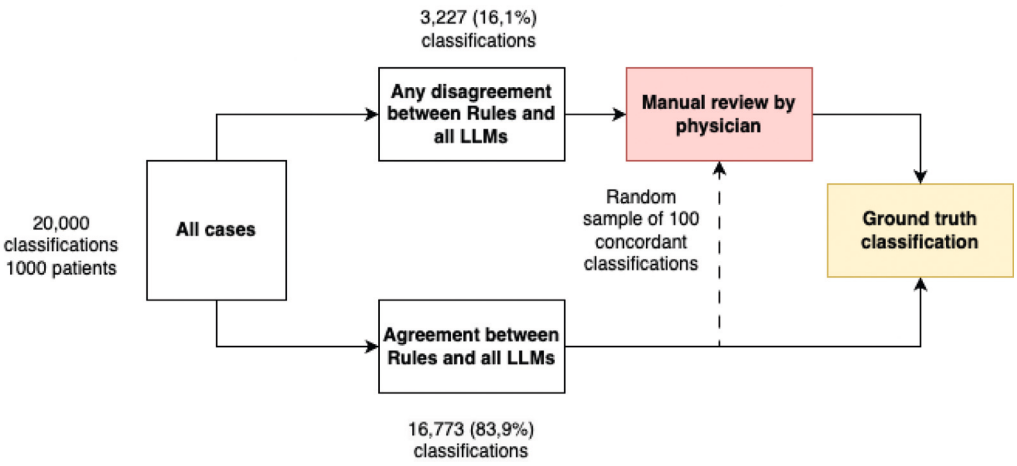


Fig. 4. Annotation process used in this study: all cases were evaluated by both LLMs and rules dictionary. All disagreements between any of the methods employed for phenotyping were reviewed by a physician. In addition, 100 random concordant cases were manually reviewed.

remaining procedure, measurement, and drug-related mappings (See Appendix).

For each patient, we evaluated the presence of all 20 chronic conditions using three approaches: LLM-based classification, dictionary rule-based classification, and condition code presence alone. Evaluation was performed using the *scikit-learn* library to calculate classification metrics [33]. The metrics reported include accuracy, precision, recall, and the F1-score. We specifically highlight the F1-score for the 'True' class, which is calculated as the harmonic mean of precision and recall for the positive (presence of a condition) class. Additionally, the macro F1-score, which averages the F1-scores across both classes, is reported to provide a balanced view of the model's overall performance.

3. Results

3.1. Overall description of the population and condition prevalences

The evaluation dataset includes 1,000 unique patients, 58.6% of whom are women, with a mean age of 56.8 years. Fig. 5 shows the total number of chronic conditions identified by each phenotyping method. There is a significant discrepancy in the number of conditions identified, with Llama 3.8b identifying the most and condition codes identifying the fewest. Fig. 6 illustrates the number of patients identified with each condition, depending on the phenotyping method used.

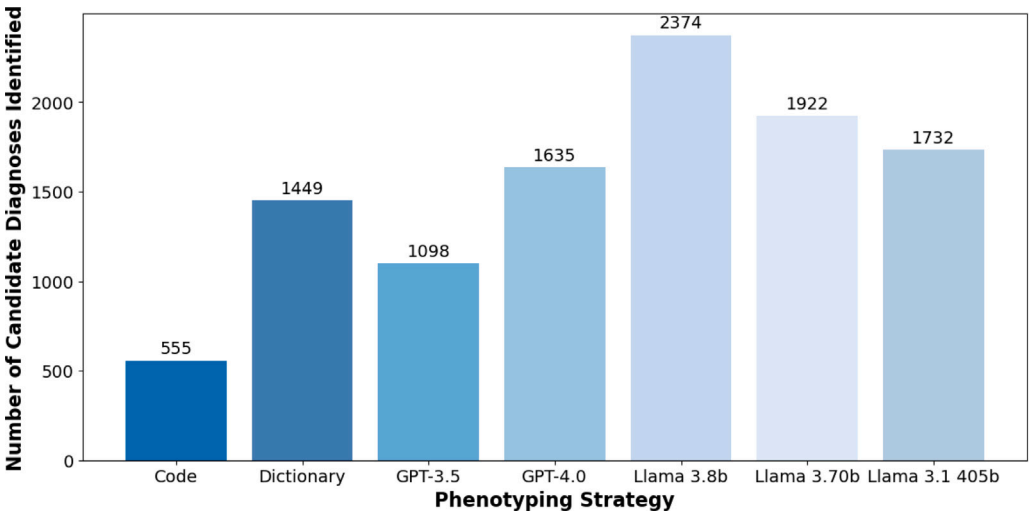


Fig. 5. Overall number of diagnoses identified among the different phenotyping strategies: condition codes, dictionary rules and LLMs.

Table 1
Prevalence of the 20 studied chronic conditions by the seven phenotyping strategies (%).

Condition	Phenotyping method						
	Condition codes (%)	Dictionary rules (%)	GPT-3.5 (%)	GPT-4o (%)	Llama 3 8b (%)	Llama 3 70b (%)	Llama 3.1 405b (%)
Hyperlipidemia	0.0	22.9	15.3	24.1	22.8	23.9	23.2
Autoimmune diseases	0.7	1.7	0.8	3.1	2.6	2.9	2.5
Hypertension	10.4	13.7	9.6	16.7	11.8	13.3	13.4
Benign prostate hypertrophy	0.0	0.0	5.4	8.2	7.0	7.6	7.6
Cancer	7.9	8.1	7.3	9.3	8.5	8.5	8.4
Heart Failure	1.5	3.2	1.5	2.8	1.8	2.3	2.4
Arrhythmias	0.0	0.0	2.8	5.0	3.9	4.2	4.5
Anxiety disorder	0.7	4.1	1.9	8.6	3.7	6.7	7.3
Disorders of thyroid gland	4.8	6.2	5.0	8.0	6.1	7.1	7.0
Ischemic heart disease	0.8	1.2	1.2	1.7	1.3	1.5	1.5
Osteoporosis	0.1	1.1	0.2	1.2	1.0	1.0	0.9
Disorder of vertebral column	5.0	5.7	7.0	14.3	11.1	12.1	11.1
Depression	1.7	3.6	1.6	4.4	2.7	3.8	3.8
Urinary incontinence	1.8	2.6	1.6	2.5	1.6	2.0	2.3
Osteoarthritis	4.7	4.7	4.2	5.3	4.1	4.5	4.6
Diabetes	3.4	4.6	4.6	6.8	3.8	6.0	6.3
Asthma	3.3	3.3	3.7	4.2	3.4	3.8	3.9
Anemia	2.4	4.4	3.4	11.1	4.4	7.4	8.0
Peripheral neuropathy	1.0	1.0	0.1	0.7	0.3	0.6	0.8
Epilepsy	0.7	0.8	0.6	1.0	0.8	0.7	0.8

As shown in Table 1, the prevalence of the 20 chronic conditions varies widely based on the phenotyping method considered. For instance, no patients with hyperlipidemia were identified using condition codes; however, dictionary rules suggest a prevalence of 22.9%, and AI models identified prevalences ranging from 15.3% to 24.1%. Some conditions, such as hypertension and cancer, exhibited more consistent prevalences across methods. Hypertension was identified in 10.4% of patients using condition codes, with AI models and dictionary rules detecting prevalences between 9.6% and 16.7%. Cancer prevalence varied from 7.9% using condition codes to 9.3% with GPT-4o. Conditions like osteoporosis and peripheral neuropathy showed consistently low prevalence rates across all methods.

3.2. Agreement rates between phenotyping methods

Overall agreement among the methods remains very high. The comparison of different phenotyping methods reveals strong concordance in identifying conditions, with most methods showing match percentages above 90% (See Table 2). The highest agreement was observed between GPT-4o and Llama 3.1 405b (96.16%), while comparisons between GPT-4o and other models also showed similarly high agreement. Conversely, the dictionary-based method and Llama 3 8b had the lowest agreement, with a disagreement rate of 10.29%. Across all phenotyping methods evaluated, there was at least 1 disagreement in 3227 (16.13%) of the classifications.

3.3. Phenotyping accuracy

The evaluation of phenotyping accuracy across different methods highlights GPT-4o as the most effective approach, surpassing both dictionary rules and all the other LLMs in both recall and macro F1-score, as shown in Table 3 and Figs. 8, 9, 10, and 7. GPT-4o demonstrated the highest performance with a precision score of 0.87, a recall score of 0.97, and an F1 score of 0.92. In contrast, condition codes, while achieving the highest precision score (0.92), significantly underperformed in recall (0.36), leading to a much lower F1 score of 0.51. This suggests that although condition codes are precise in identifying negative cases, they struggle to detect true positives, especially in conditions like hyperlipidemia and osteoporosis where very few patients were identified.

The Dictionary Rules method achieved a moderate balance, with a precision of 0.65, a recall of 0.65, and an F1 score of 0.65. Among the remaining LLMs, GPT-3.5 and the Llama models exhibited varying levels of effectiveness. GPT-3.5 had a precision of 0.72, a recall of 0.54,

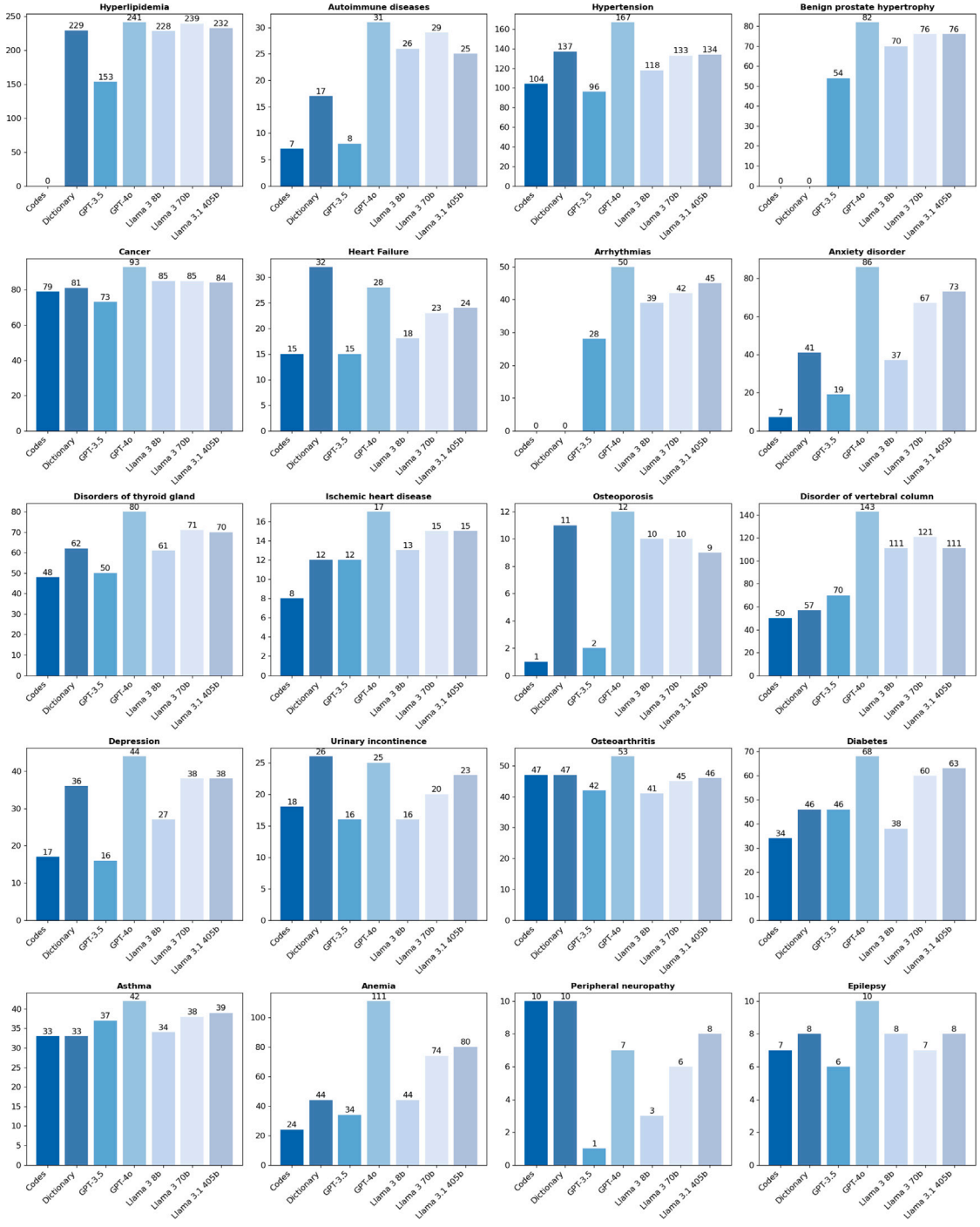


Fig. 6. Number of candidate diagnoses made per condition, depending on phenotyping strategy.

and an F1 score of 0.62, slightly outperforming Llama 3 8b, which had the lowest F1 score (0.55) among all evaluated LLMs. Llama 3 70b and Llama 3.1 405b showed improved performance, with higher recall and better precision, resulting in F1 scores of 0.72 and 0.77, respectively.

The complete counts of false positive and false negative classifications per method can be found in Fig. 8. Examples of situations where each method failed to identify a positive case or wrongly assigned a condition where true label was found are depicted in Tables 5 and 6.

3.4. AI models confidence in classification

All models demonstrated improved classification performance with increasing confidence levels, effectively reflecting the degree of confidence expressed by the model in their predictions (See Table 4). At the high confidence level, GPT-4o achieved near-perfect accuracy (0.989), precision (0.949), and recall (0.983), handling the majority of cases (16,186) with strong reliability. Even at the medium confidence level, GPT-4o maintained robust performance, with an accuracy of 0.972 and recall of 0.856, though precision slightly decreased to 0.774.

Table 2
Agreement comparison between phenotyping strategies.

Comparison	Agreements (n, %)	Disagreements (n, %)
Rules dictionary vs GPT-3.5	18,697 (93.48%)	1303 (6.52%)
Rules dictionary vs GPT-4o	18,754 (93.77%)	1246 (6.23%)
Rules dictionary vs Llama 3 8b	17,942 (89.71%)	2058 (10.29%)
Rules dictionary vs Llama 3 70b	18,471 (92.36%)	1529 (7.64%)
Rules dictionary vs Llama 3.1 405b	18,615 (93.08%)	1385 (6.92%)
GPT-3.5 vs GPT-4o	18,891 (94.45%)	1109 (5.55%)
GPT-3.5 vs Llama 3 8b	18,192 (90.96%)	1808 (9.04%)
GPT-3.5 vs Llama 3 70b	18,668 (93.34%)	1332 (6.66%)
GPT-3.5 vs Llama 3.1 405b	18,766 (93.83%)	1234 (6.17%)
GPT-4o vs Llama 3 8b	18,262 (91.31%)	1738 (8.69%)
GPT-4o vs Llama 3 70b	19,109 (95.55%)	891 (4.45%)
GPT-4o vs Llama 3.1 405b	19,231 (96.16%)	769 (3.84%)
Llama 3 8b vs Llama 3 70b	18,417 (92.08%)	1583 (7.92%)
Llama 3 8b vs Llama 3.1 405b	18,280 (91.40%)	1720 (8.60%)
Llama 3 70b vs Llama 3.1 405b	19,170 (95.85%)	830 (4.15%)

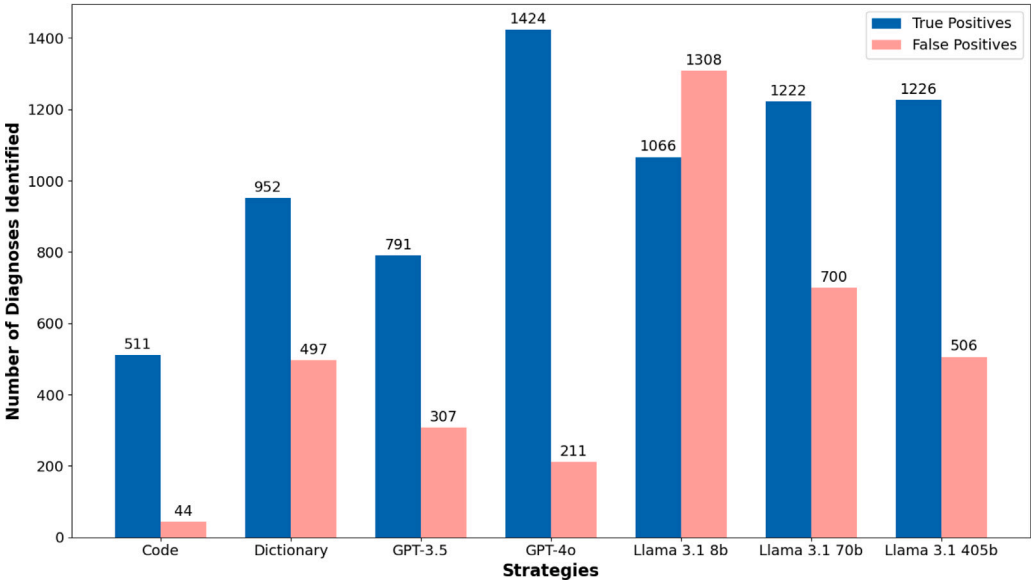


Fig. 7. Number of true and false positives obtained with the evaluated phenotyping strategies.

Table 3
Overall classification metrics for the seven evaluated phenotyping strategies.

Model	Accuracy	Precision	Recall	F1-score	Macro F1-score
Condition codes	0.95	**0.92**	0.36	0.51	0.74
Rules dictionary	0.95	0.65	0.65	0.65	0.81
GPT-3.5	0.95	0.72	0.54	0.62	0.80
GPT-4o	**0.99**	0.87	**0.97**	**0.92**	**0.96**
Llama 3 8b	0.92	0.44	0.73	0.55	0.75
Llama 3 70b	0.95	0.63	0.84	0.72	0.85
Llama 3.1 405b	0.96	0.70	0.84	0.77	0.87

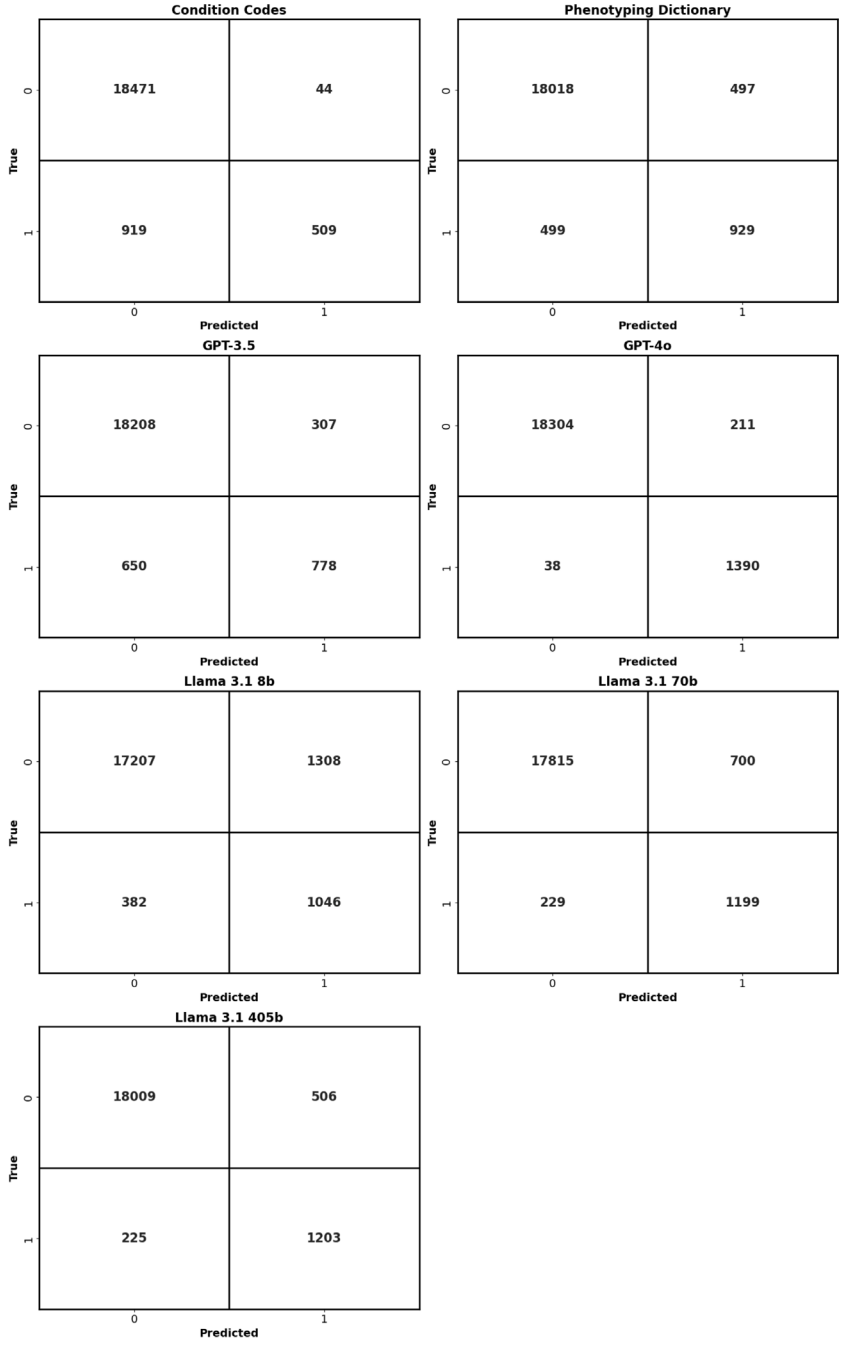
GPT-3.5, while retaining high precision (0.862) at the high confidence level, exhibited lower recall (0.762), indicating some challenges in consistently identifying true positives. The Llama models displayed varying effectiveness. Llama 3 70b and Llama 3.1 405b both performed well at high confidence, but still below GPT-4o, with F1 scores of 0.903 and 0.933, respectively. Llama 3 8b handled a large portion of classifications at low confidence, resulting in a low F1 score (0.504). Almost all models struggled at low confidence levels, with significant drops in performance across the board.

Table 4
Classification performance of the five LLMs across confidence levels.

Model	Confidence level	Accuracy	Precision	Recall	F1 score	Count
GPT-3.5	Low	0.961	0.481	0.500	0.490	5015
	Medium	0.836	0.792	0.807	0.799	1746
	High	0.961	0.862	0.762	0.803	13,239
GPT-4o	Low	1.000	1.000	1.000	1.000	495
	Medium	0.972	0.774	0.856	0.809	3319
	High	0.989	0.949	0.983	0.966	16,186
Llama 3 8b	Low	0.966	0.505	0.504	0.504	13,535
	Medium	0.423	0.553	0.575	0.415	1318
	High	0.902	0.828	0.898	0.855	5146
Llama 3 70b	Low	0.950	0.809	0.539	0.559	482
	Medium	0.647	0.617	0.679	0.598	1401
	High	0.975	0.887	0.922	0.903	18,117
Llama 3.1 405b	Low	0.959	0.581	0.508	0.508	2574
	Medium	0.840	0.705	0.813	0.736	2388
	High	0.982	0.915	0.952	0.933	15,038

3.5. Explanations provided by LLMs for phenotyping classification

Another dimension evaluated in this pipeline is the justification for a decision provided by the model. In all inferences, LLMs wrote



tot

Fig. 8. Confusion matrices on phenotyping classification among the seven evaluated strategies.

a rationale behind the decision if classifying the presence of a given condition. This rationale was helpful during annotation process and enabled confirmation of the decisions made on each inference. Table 5 presents selected cases exemplifying the reasoning capabilities and limitations of LLMs in phenotyping chronic conditions. These examples demonstrate how LLMs can effectively synthesize diverse clinical information, such as medication history, diagnostic procedures, and clinical notes, to infer the presence of chronic conditions even when they are not explicitly mentioned. For instance, in the case of asthma, GPT-4o integrates multiple data points, including medication history and clinical diagnoses, to confirm the diagnosis. Meanwhile, the hyperlipidemia case highlights GPT-4o's capacity to critically evaluate clinical data, avoiding false positives by correctly assessing that the patient's cholesterol levels do not indicate hyperlipidemia. Finally, the cancer case reveals how GPT-4o can infer diagnoses from specific surgical procedures, making it more reliable in identifying certain chronic conditions compared to GPT-3.5.

3.6. Evaluation of fairness and bias across demographics

The evaluation of fairness in chronic condition phenotyping methods using condition codes, dictionaries, and LLMs reveals minimal gender bias but highlights noticeable age-related disparities, as detailed in Tables 7 and 8. Across genders, all methods demonstrate balanced performance in terms of Accuracy, Precision, Recall, and F1 Score, with

Table 5
Examples of false positive and false negative LLMs classifications and rationale.

Method	Outcome	LLM Rationale	Comments
LLMs			
GPT-3.5	False Positive	"The patient has undergone multiple procedures and measurements related to cardiovascular health, such as Doppler ultrasonography of carotid artery, echocardiography, and blood pressure measurements. These procedures and measurements are commonly used to diagnose and monitor hypertension".	No explicit information of existence of hypertension.
GPT-3.5	False Negative	"There is no specific evidence in the clinical note to suggest the presence of anemia".	Anemia was indeed identified in laboratory measurements.
GPT-4o	False Positive	"The patient was prescribed captopril, an antihypertensive medication, at age 32, indicating a diagnosis of hypertension".	Erroneously inferred hypertension despite insufficient corroborating evidence.
GPT-4o	False Negative	"There is no mention of cancer or related terms such as malignancy or tumor in the clinical note".	Failed to detect prostate cancer as detected by performance of radical prostatectomy with lymphadenectomy.
Llama 3 8b	False Positive	"The patient has a diagnosis of Myelosclerosis with myeloid metaplasia, which is a type of cancer".	Annotator and dictionary rules did not consider this entity as a cancer.
Llama 3 8b	False Negative	"The patient's clinical note does not show any measurements of blood pressure, which is a common indicator of hypertension".	Missed hypertension despite evidence of losartan prescriptions.
Llama 3 70b	False Positive	"The patient underwent multiple MRIs and X-rays of the knee, which could be indicative of osteoarthritis".	Osteoarthritis diagnosis cannot be inferred solely from this information.
Llama 3 70b	False Negative	"The patient has a condition listed as 'Shoulder joint pain' and has undergone a plain X-ray of the shoulder, but there is no evidence of a disorder of the vertebral column".	Failed to identify a patient who underwent imaging of the lumbar spine and had conditions like 'Pain in thoracic spine'.
Llama 3 405b	False Positive	"The patient has high cholesterol levels (169.0, 189.0) and high triglyceride levels (96.0, 110.0) in multiple measurements, indicating hyperlipidemia".	Hyperlipidemia not confirmed through phenotyping definition.
Llama 3 405b	False Negative	"There is no indication of urinary incontinence in the clinical note"	Failed to identify "Female stress urinary incontinence" explicit mention.

Table 6
Examples of false positive and false negative condition codes and dictionary entries classifications.

Method	Outcome	Comments
Condition Codes		
Condition Code	False Positive	Benign neoplasm of skin of trunk code erroneously assigned to cancer.
Condition Code	False Negative	Missed condition code for hypertension despite evidence of anti-hypertensive treatment.
Dictionary Entries		
Dictionary Rules	False Positive	Dictionary definition erroneously triggered Inflammatory dermatosis as an Autoimmune Disease.
Dictionary Entry	False Negative	Dictionary definition failed to capture a patient that showed several "Low back pain" and "MRI of lumbar spine" codes.

only minor variations. However, the analysis by age group shows a consistent decline in classification performance as age increases (Table 8). All approaches exhibit higher F1 scores in the 18–30 age group, with the lowest scores observed in the 90–100 age group (see Fig. 11). GPT-4o stands out as the model with the highest F1 scores across all age groups, while also being the most robust to bias, displaying the

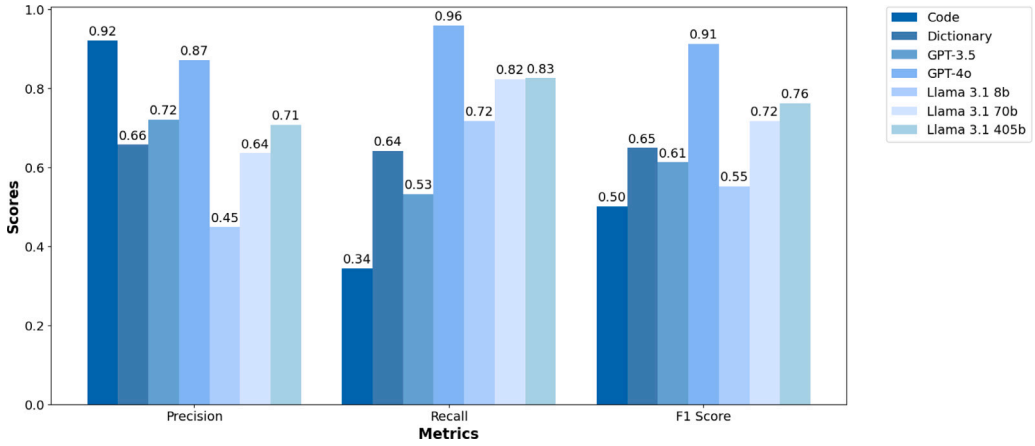


Fig. 9. Overall precision, recall and F1-score for the seven evaluated phenotyping strategies.

Table 7
Classification performance by gender across the seven phenotyping strategies.

Method	Gender	Accuracy	Precision	Recall	F1 score	Count
Condition codes	Female	0.952	0.950	0.952	0.942	11,720
	Male	0.946	0.945	0.946	0.933	8280
Rules dictionary	Female	0.952	0.953	0.952	0.952	11,720
	Male	0.944	0.941	0.944	0.942	8280
GPT-3.5	Female	0.950	0.945	0.950	0.946	11,720
	Male	0.950	0.945	0.950	0.947	8280
GPT-4o	Female	0.986	0.987	0.986	0.987	11,720
	Male	0.986	0.987	0.986	0.987	8280
Llama 3 8b	Female	0.910	0.936	0.910	0.920	11,720
	Male	0.918	0.939	0.918	0.926	8280
Llama 3 70b	Female	0.951	0.959	0.951	0.954	11,720
	Male	0.953	0.961	0.953	0.955	8280
Llama 3.1 405b	Female	0.962	0.965	0.962	0.963	11,720
	Male	0.962	0.966	0.962	0.964	8280

Table 8
Classification performance by age group across the seven phenotyping strategies.

Method	Age group	Accuracy	Precision	Recall	F1 score	Count
Condition codes	18–30	0.975	0.972	0.975	0.969	1180
	30–40	0.966	0.962	0.966	0.959	2360
	40–50	0.963	0.959	0.963	0.955	3740
	50–60	0.957	0.955	0.957	0.949	3760
	60–70	0.942	0.941	0.942	0.929	3740
	70–80	0.930	0.932	0.930	0.915	2880
	80–90	0.922	0.928	0.922	0.903	1980
	90–100	0.908	0.912	0.908	0.891	360
Rules dictionary	18–30	0.961	0.966	0.961	0.963	1180
	30–40	0.961	0.964	0.961	0.962	2360
	40–50	0.964	0.966	0.964	0.965	3740
	50–60	0.956	0.958	0.956	0.956	3760
	60–70	0.945	0.942	0.945	0.943	3740
	70–80	0.927	0.927	0.927	0.927	2880
	80–90	0.927	0.921	0.927	0.923	1980
	90–100	0.917	0.911	0.917	0.909	360
GPT-3.5	18–30	0.981	0.980	0.981	0.981	1180
	30–40	0.967	0.963	0.967	0.964	2360
	40–50	0.966	0.963	0.966	0.964	3740
	50–60	0.963	0.960	0.963	0.961	3760
	60–70	0.944	0.939	0.944	0.940	3740
	70–80	0.929	0.923	0.929	0.924	2880
	80–90	0.908	0.895	0.908	0.897	1980
	90–100	0.897	0.890	0.897	0.879	360

(*continued on next page*)

Table 8 (*continued*).

Method	Age group	Accuracy	Precision	Recall	F1 score	Count
GPT-4o	18–30	0.992	0.993	0.992	0.993	1180
	30–40	0.990	0.991	0.990	0.991	2360
	40–50	0.987	0.988	0.987	0.987	3740
	50–60	0.991	0.991	0.991	0.991	3760
	60–70	0.984	0.986	0.984	0.985	3740
	70–80	0.984	0.985	0.984	0.984	2880
	80–90	0.979	0.980	0.979	0.980	1980
	90–100	0.969	0.971	0.969	0.970	360
Llama 3 8b	18–30	0.940	0.975	0.940	0.953	1180
	30–40	0.939	0.960	0.939	0.947	2360
	40–50	0.924	0.959	0.924	0.936	3740
	50–60	0.920	0.947	0.920	0.930	3760
	60–70	0.899	0.930	0.899	0.910	3740
	70–80	0.905	0.915	0.905	0.909	2880
	80–90	0.886	0.896	0.886	0.890	1980
	90–100	0.867	0.869	0.867	0.868	360
Llama 3 70b	18–30	0.977	0.983	0.977	0.979	1180
	30–40	0.966	0.971	0.966	0.968	2360
	40–50	0.964	0.973	0.964	0.967	3740
	50–60	0.959	0.967	0.959	0.961	3760
	60–70	0.944	0.956	0.944	0.948	3740
	70–80	0.938	0.942	0.938	0.939	2880
	80–90	0.930	0.937	0.930	0.933	1980
	90–100	0.889	0.906	0.889	0.895	360
Llama 3.1 405b	18–30	0.982	0.985	0.982	0.983	1180
	30–40	0.972	0.976	0.972	0.974	2360
	40–50	0.975	0.978	0.975	0.976	3740
	50–60	0.967	0.971	0.967	0.969	3760
	60–70	0.960	0.964	0.960	0.961	3740
	70–80	0.943	0.945	0.943	0.944	2880
	80–90	0.939	0.944	0.939	0.941	1980
	90–100	0.922	0.927	0.922	0.924	360

smallest variation in performance (−0.023) across age groups, even outperforming condition coding and the rules dictionary in maintaining consistent classification performance.

4. Discussion

In this study, we applied zero-shot learning using LLMs on EHR data, demonstrating that this approach can significantly enhance traditional rule-based phenotyping, even in the absence of extensive labeled datasets. Accurate phenotyping of chronic conditions is especially crucial in the context of multimorbidity, where patients often suffer from multiple coexisting diseases, complicating both diagnosis and treatment. By improving the identification of these conditions, our approach can not only enhance individual patient care but also facilitate more effective management of complex patient populations.

Fig. 10. Classification metrics of each phenotyping strategy across all conditions.

Among the methods evaluated, GPT-4o outperformed both GPT-3.5 and all the Llama 3 models, as well as the rule-based phenotyping algorithm across all conditions. These results are in line with prior findings where GPT-4 demonstrated superior performance in identifying clinical phenotypes from EHR data [34]. Moreover, GPT-4o exhibited excellent calibration across gender and age strata, maintaining very high performance when its confidence was high and still delivering robust results at moderate confidence levels. This improved ability to capture both direct and indirect condition mentions not only increases the number of correctly identified patients but also aligns with recent work highlighting the efficiency of LLMs in reducing manual annotation efforts [19,20]. Our approach also relates to previous studies that employ ensemble learning and active learning techniques to improve information extraction from EHRs [23,35]. By leveraging multiple strategies and concentrating efforts on discordant cases, our approach optimizes the accuracy of phenotyping while efficiently allocating manual annotation resources to the most challenging cases. This targeted approach not only refines phenotyping accuracy but also facilitates the continuous improvement of both rule-based methods and LLMs, ensuring they remain effective in real-world clinical settings.

Another important aspect of our findings is the ability of AI models to accurately assess and communicate their confidence in predictions. This study demonstrated that when these models express high confidence, their predictions are not only highly accurate but also reliable across various conditions. Even at moderate confidence levels, GPT models and Llama 3.1 405b maintain reasonable performance,

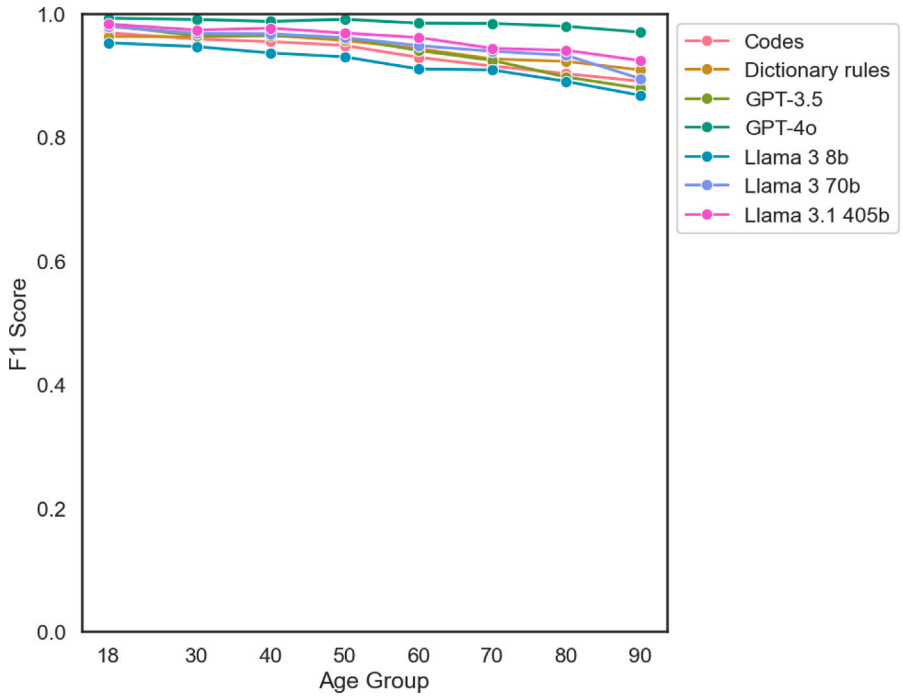


Fig. 11. F1-score across different age strata for the seven evaluated phenotyping strategies.

underscoring their robustness. Interestingly, GPT-4o also exhibited a conservative approach in low-confidence cases, where all predictions were correct negatives, indicating its reliability in minimizing false positives when uncertainty is high. These findings are in line with only a few recent studies that evaluate model confidence on answers through prompting techniques which show a correlation between larger models (such as GPT-4o) confidence and accuracy [34,36]. However, even if LLMs can provide detailed rationales that enhance transparency and interpretability, their internal confidence in predictions may not necessarily align with human-perceived confidence, motivating current research on ensuring their reliability in clinical settings [37].

We investigated the presence of bias across gender and different age groups in this study and consistently observed a decline in performance among older age groups across all methods. Previous studies have highlighted the risks of LLMs introducing and propagating bias in healthcare setting [38]. In our study, LLMs for phenotyping conditions actually resulted in reduced bias compared to standard coding and rule-based phenotyping. Ensuring fairness in AI-driven tools is crucial, and our findings suggest that advanced models like GPT-4o are making progress in this area. However, this issue requires further exploration in larger and more diverse datasets to ensure that AI systems deliver equitable and reliable outcomes across all populations, making continued research in this area a top priority.

Despite these strengths, our study has several limitations. First, developing effective prompts was time-consuming and required iterative testing to achieve reliable results. Early in the process, we encountered issues such as hallucinations, where the models generated incorrect or nonsensical information. Additionally, enforcing a consistent JSON output in these models proved challenging, particularly in GPT-3.5 and Llama 3 7b. Although prompt engineering techniques helped mitigate these issues, the process was largely empirical and lacked a systematic approach. Developing a more structured methodology for prompt creation could improve efficiency and consistency in results, reducing the trial-and-error nature of prompt refinement.

Another limitation is the use of synthetic summaries generated from text descriptions of structured codes, which differ from actual clinical notes. These summaries often lacked context or contained ambiguous information, complicating the determination of ground truth. Real-world clinical notes are typically more complex, with jargon, abbreviations, and inconsistencies that would likely require multiple annotators or iterative rounds of annotation to ensure accuracy. Consequently, our results may not fully reflect the challenges of applying LLMs to real-world data, and further research is needed to assess model performance on more representative datasets.

Furthermore, the models used in this study were not specifically trained on medical data, unlike specialized models like MedPalm [39]. Additionally, models with larger context windows, such as Anthropic's Claude, could potentially improve phenotyping predictions by capturing more longitudinal clinical information. While we recognize the value of testing these more specialized models, they were not readily accessible within our data infrastructure.

Our validation method is robust, given that we compare different phenotyping methods and use manual annotation for disagreement cases. However, this assumes that dictionary and code definitions are comprehensive and accurate. Although OMOP vocabularies provide a comprehensive mapping strategy for clinical codes, it is known that they have limitations in specific scenarios, such as Oncology and even Primary Care [40,41]. Another limitation concerns the way we developed our dictionary-based phenotyping rules. Although there was a laborious curation effort in defining these rules, we ultimately found additional ambiguous cases when manually reviewing cases against LLM predictions. This led to dictionary false positives, such as a single prescription of a short-acting drug for hypertension (*e.g.*, captopril) or benign neoplasms being mapped to cancer (See Table 6). While this is an obvious limitation of our rules dictionary, it also highlights the potential of our LLM-based approach to further refine phenotyping rules over time.

Finally, while GPT-4o outperformed other methods in phenotyping accuracy, this study did not analyze the cost implications of scaling this approach to an entire EHR database. The sample size of 1,000 patients used in this study was sufficient for demonstration purposes, but larger-scale implementations would need to consider the computational resources and potential costs associated with deploying such advanced models. Additionally, while GPT-4o was the most effective model evaluated, it should be noted that we only evaluated large models trained on general-domain data. The rapid evolution of LLMs

Table A.9
Phenotyping rules dictionary. The table provides a comprehensive mapping of conditions, procedures, laboratory measurements, and drug prescriptions used for phenotyping. It is based on standardized vocabularies from Observational Health Data Sciences and Informatics (OHDSI), ensuring that each condition code and procedure code (standard OMOP concept id) represents an ancestor concept, with all hierarchical descendants automatically included. Drug prescriptions are mapped using an adapted version of the Anatomical Therapeutic Chemical (ATC) classification system. Additionally, specific conditions such as heart failure, diabetes mellitus, hyperlipidemia, and anemia are identified using laboratory measurements based on standard definitions.

Condition	Concepts
Coronary Artery Disease	Condition: Ischemic heart disease (4185932). Drugs: C07AA01–C07AA06, C07AA08–C07AB01, C07AB02, C07AG01 C08CA01–C08DB01, C09DB01–C09DB04, C09DX01, C09BB02–C09BB10, C07AB03 C09DX03, C10BX03, C01DA02–C01DA14, C01DX16, C08EX02; Procedures: Procedure on coronary arteries (4043174)
Aortic aneurysm	Condition: Aortic aneurysm (317585). Procedures: Repair of aneurysm of aorta (4049675)
Arrhythmias	Condition: Arrhythmia (44784217). Drugs: C01AA05 C01BA01–C01BD01 C07AA07; Procedures: Cardioverter defibrillator procedure (4138751); Automatic cardiac defibrillator (4232661); Cardiac pacemaker procedure (4051938); Procedure for arrhythmia (4051932)
Heart failure	Condition: Congestive Heart Failure (316139). Drugs: C03DA02–C03DA99 C07AB07 C07AG02 C07AB12; C09DX04 Measurements: pro-BNP (3029187); Criterion: > 125 pg/ml
Hypertension	Condition: Hypertensive disorder (316866); Drugs: C03AA01–C03BA11 C03DB01 C03DB99 C03EA01 C09BA02–C09BA09 C09DA02-C09DA08 C02AB01–C02AC05 C02DB02-C02DB99
Valvular heart disease	Condition: Valvular heart disorder (4281749); Procedures: Procedure on heart valve (4042674)
Peripheral artery disease	Condition: Peripheral vascular disease (321052); Procedures: Angioplasty of peripheral blood vessel (46271049)
Diabetes mellitus	Condition: Diabetes mellitus (201820); Drugs: A10AA01–A10BX99; Measurements: HbA1c (3005673); Criterion: >6.5%
Hyperlipidemia	Condition: Hyperlipidemia (4031814); Drugs: C10AA01–C10BX0; Measurements: Total cholesterol (3027114); Criterion: > 190
Disorders of thyroid gland	Condition: Disorders of thyroid gland (141253); Drugs: H03AA01–H03AA02 H03BA02 H03BB01; Procedures: Radionuclide therapy for hyperthyroidism (4275582); Biopsy of thyroid (4085189)
Obesity	Condition: Obese (4215968); Drugs: A08AA01- A08AX01; Procedures: Bariatric operative procedure (4326683)
Chronic Liver Disease	Condition: Chronic Liver Disease (4212540); Drugs: A07AA11; Procedures: Operation on esophageal varices (4122243)
Gallstones	Condition: Gallstones (4145627); Procedures: Cholecystectomy (4242997) Bile duct calculus procedure (4341100)
Chronic pancreatitis	Condition: Chronic pancreatitis (195596); Drugs: A09AA02
Upper digestivex tract disease	Condition: Disorder of upper gastrointestinal tract (4000609) — excluding all existing codes elsewhere; Drugs: A02BA01–A02BX05
Chronic bowel disease	Condition: Disorder of lower gastrointestinal tract (4197094) — excluding all existing codes elsewhere; Drugs: A07EC01–A07EC04 A07EA01–A07EA02 A07EA06 L04AA33
Anemia Chronic	Condition: Anemia (439777) Procedures: Intravenous infusion of iron (36713609); Transfusion of red blood cells (4022173) Measurements: Hb (3000963); Criterion: <12g/dl women < 13 g/dl men in 2 measurements with at least 6 months apart.
HIV infection	Condition: HIV (439727); Drugs: J05AE01–J05AE10 J05AF12–J05AG05 J05AR01–J05AR99 J05AX07–J05AX09 J05AX12 J05AF01–J05AF07 J05AF09;
Chronic skin condition	Condition: Chronic disease of the skin (4134132) Drugs: D05AA01–D05AA99 D05BB01 D05BB02 D05AX02 D05AC01–D05AC51 D05AX52
Autoimmune diseases	Condition: Autoimmune diseases (434621)
Gout	Condition: Gout (440674) Drugs: M04AA01–M04AC01
Osteoarthritis	Condition: Osteoarthritis (80180);

(*continued on next page*)

Table A.9 (*continued*).

Condition	Concepts
Osteoporosis	Condition: Osteoporosis (80502) Drugs: M05BA01–M05BB05 M05BX03 M05BX04 G03XC01 H05AA02
Disorder of vertebral column	Condition: Disorder of vertebral column (44782549) Procedures: Operation on intervertebral disk (4034458); Procedure on vertebra (4040536)
Dementia	Condition: Dementia (4182210) Drugs: N06DA02–N06DA04 N06DX01
Migraine	Condition: Migraine (318736) Drugs: N02CA01–N02CX01
Multiple sclerosis	Condition: Multiple sclerosis (374919)
Parkinson's disease	Condition: Parkinson's disease (4140090) Drugs: N04AA01–N04BX02
Epilepsy	Condition: Seizure disorder (4029498) Drugs: N03AA01–N03AX99
Cerebrovascular disease (including stroke)	Condition: Cerebrovascular disease (381591)
Paralysis	Condition: Paralysis (440377)
Peripheral neuropathy	Condition: Peripheral neuropathy (4117779)
Sleep disorders	Condition: Sleep disorders (435524)
Chronic pain	Condition: Chronic pain (436096); Drugs: N02AA01–N02AX02 N02AX06 N02AX52 N02BE51 – If prescribed in 2 or more episodes in at least 2 consecutive years
Cancer	Condition: Malignancies (443392); Drugs: L01AA01–L01XX41; Procedures: Radiation oncology and/or therapy (4029715); Chemotherapy (4273629)
Visual impairment	Condition: Visual impairment and/or blindness (4023310); Cataract (375545); Procedures: Cataract surgery (4004519)
Hearing loss	Condition: Hearing loss (377889); Procedures: Deafness remedial therapy (4084061)
Alcohol use disorders	Condition: Alcoholism (4218106); Drugs: N07BB01–N07BB99
Drug use disorders	Condition: Substance dependence (37165431)
Psychotic disorder	Condition: Psychotic disorder (436073); Drugs: N05AA01–N05AB02 N05AB06–N05AL07 N05AX07–N05AX13
Anxiety disorder	Condition: Anxiety disorder (442077); Drugs: N05BA01–N05BA12 N05BE01
Nicotine dependency	Condition: Nicotine dependency (4209423) Drugs: N07BA01–N07BA03 N06AX12
Bipolar disorder	Condition: Bipolar disorder (436665) Drugs: N05AN01
Depression	Condition: Depression (440383) Drugs: N06AA01–N06AG02 N06AX03–N06AX11 N06AX13–N06AX18 N06AX21–N06AX26
Asthma	Condition: Asthma (317009) Drugs: R03DC-R03DX
COPD	Condition: COPD (255573) Drugs: R03AC02– R03DB06
Pulmonary fibrosis	Condition: Fibrosis of lung (4197819)
Chronic kidney disease	Condition: Chronic kidney disease (46271022) Drugs: B03XA01–B03XA03 A11CC01–A11CC04 V03AE02 V03AE03 V03AE05 Measurement: Serum creatinine > 12mg/dl (women) > 14 mg/dl men Procedures: Renal dialysis (4146536)
Fluid and electrolyte disorders	Condition: Disorder of fluid AND/OR electrolyte (441830) Drugs: V03AE01
Urinary incontinence	Condition: Urinary incontinence (197672) Drugs: G04BD01–G04BD99 Procedures: Repair of urinary stress incontinence (4242750)
Urolithiasis	Condition: Urolithiasis (4319447); Procedures: Lithotripsy (4046290); Urinary calculus removal (4170611); Insertion of stent into ureter (4181450)
Benign prostate hypertrophy	Condition: Benign prostate hypertrophy (197032); Drugs: G04CA01–G04CA99 G04CB01 G04CB02 — if male gender Procedure: Transurethral prostatectomy (4234536)
Endometriosis	Condition: Endometriosis (433527)

(*continued on next page*)

Table A.9 (*continued*).

Condition	Concepts
Uterine fibroid	Condition: Uterine fibroid (37016195) Procedures: Uterine myomectomy (4169931)
Malnutrition	Condition: Malnutrition (435227) Drugs: B05BA01–B05BA10

suggests that smaller, specialized models trained on biomedical data might offer comparable or even superior performance at a lower cost. Comparative analysis of these models should be a focus of future research.

5. Conclusion

In conclusion, we propose a phenotyping pipeline that leverages the integration of LLMs with traditional rule-based methods, significantly reducing the need for extensive manual data annotation, streamlining the phenotyping process, and reducing dependency on human resources while providing transparent prediction confidence and rationales. Our study demonstrates that zero-shot learning with large language models (LLMs), particularly GPT-4o, offers a powerful and efficient approach for phenotyping chronic conditions from EHRs, by accurately identifying a wide range of chronic conditions without model training and using minimal labeling effort.

To fully harness the potential of LLMs in healthcare, future research should focus on refining prompt engineering to enhance model efficiency and consistency, improving model performance on real-world clinical data, and exploring cost-effective strategies for broader implementation. Addressing these challenges will pave the way for more accurate and widely applicable phenotyping solutions, ultimately advancing clinical research and improving patient care.

CRediT authorship contribution statement

Bernardo Neves: Writing – review & editing, Writing – original draft, Visualization, Validation, Software, Resources, Project administration, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **José Maria Moreira:** Writing – review & editing, Supervision. **Simão Gonçalves:** Writing – review & editing, Methodology. **Jorge Cerejo:** Writing – review & editing, Methodology. **Nuno A. da Silva:** Writing – review & editing, Project administration, Funding acquisition. **Francisca Leite:** Writing – review & editing, Project administration, Funding acquisition. **Mário J. Silva:** Writing – review & editing, Supervision, Conceptualization.

Ethics approval and consent to participate

This study was approved by the local Institutional Review Board from Hospital da Luz Lisboa (HLL). The ethical approval number by IRB is CES/03/2021/ME. No informed consent was collected because the project is secondary analysis of anonymized data, obtained through fully automated processes (no human interaction). Thus, the local IRB granted waivers for inform consent.

Funding

This work was supported by the IntelligentCare project LISBOA-01-0247-FEDER-045948, which is co-financed by the ERDF/LISBOA2020. Also funded by Fundação para a Ciência e a Tecnologia (FCT), under CMU-Portugal and its INESC-ID and LASIGE Research Units, ref. UIDB/50021/2020, ref. UIDB/00408/2020, and ref. UIDP/00408/2020.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper. Additionally, none of the authors have any commercial associations, financial involvements, or conflicts of interest with any entities or organizations that could be perceived to have a conflict with the content or findings of this manuscript. All authors contributed to the research independently and have no vested interest in the outcomes that would create a conflict of interest.

Acknowledgments

We thank Alexandra B. Horta, José A. Gomes, Hélder Dores, João Colaço and Gonçalo Duarte for their contribution in creating the phenotyping rules dictionary.

Appendix

See Table A.9.

Data availability

The data used in this study, while obtained from anonymized Electronic Health Records, are subject to strict confidentiality and privacy regulations. Data access requests may be considered on a case-by-case basis and will require approval from the relevant institutional review boards and data custodians. Researchers interested in obtaining access to the data for the purpose of validating or extending the findings presented in this paper should contact the corresponding author for further information on the data access process and the necessary legal and ethical requirements. All code developed for this study is shared publicly.

References

[1] M.C. Johnston, M. Crilly, C. Black, G.J. Prescott, S.W. Mercer, Defining and measuring multimorbidity: a systematic review of systematic reviews, Eur. J. Pub. Health 29 (1) (2019) 182–189, Publisher: Oxford University Press.

[2] C. Harrison, H. Britt, G. Miller, J. Henderson, Examining different measures of multimorbidity, using a large prospective cross-sectional study in Australian general practice, BMJ Open 4 (7) (2014) e004694, http://dx.doi.org/10.1136/bmjopen-2013-004694, URL https://bmjopen.bmj.com/content/4/7/e004694, Publisher: British Medical Journal Publishing Group Section: Geriatric medicine.

[3] L.E. Griffith, A. Gruneir, K.A. Fisher, K. Nicholson, D. Panjwani, C. Patterson, M. Markle-Reid, J. Ploeg, A.S. Bierman, D.B. Hogan, R. Upshur, Key factors to consider when measuring multimorbidity: Results from an expert panel and online survey, J. Comorbidity 8 (1) (2018) http://dx.doi.org/10.1177/2235042X18795306, 2235042X18795306, URL https://doi.org/10.1177/2235042X18795306, Publisher: SAGE Publications Ltd STM.

[4] I.I.S.-S. Ho, A. Azcoaga-Lorenzo, A. Akbari, C. Black, J. Davies, P. Hodgins, K. Khunti, U.T. Kadam, R.A. Lyons, C. McCowan, S.W. Mercer, K. Nirantharakumar, B. Guthrie, Examining variation in the measurement of multimorbidity in research: a systematic review of 566 studies, Lancet Public Heal. (2021) URL https://abdn.pure.elsevier.com/en/publications/examining-variation-in-the-measurement-of-multimorbidity-in-resea, Publisher: Elsevier.

[5] M. Simard, E. Rahme, A.C. Calfat, C. Sirois, Multimorbidity measures from health administrative data using ICD system codes: A systematic review, Pharmacoepidemiol. Drug Safety 31 (1) (2022) 1–12, http://dx.doi.org/10.1002/pds.5368, URL https://onlinelibrary.wiley.com/doi/abs/10.1002/pds.5368, _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/pds.5368.

[6] I.S. Kohane, B.J. Aronow, P. Avillach, B.K. Beaulieu-Jones, R. Bellazzi, R.L. Bradford, G.A. Brat, M. Cannataro, J.J. Cimino, N. García-Barrio, N. Gehlenborg, M. Ghassemi, A. Gutiérrez-Sacristán, D.A. Hanauer, J.H. Holmes, C. Hong, J.G. Klann, N.H.W. Loh, Y. Luo, K.D. Mandl, M. Daniar, J.H. Moore, S.N. Murphy, A. Neuraz, K.Y. Ngiam, G.S. Omenn, N. Palmer, L.P. Patel, M. Pedrera-Jiménez, P. Sliz, A.M. South, A.L.M. Tan, D.M. Taylor, B.W. Taylor, C. Torti, A.K. Vallejos, K.B. Wagholikar, Consortium For Clinical Characterization Of COVID-19 By EHR (4CE), G.M. Weber, T. Cai, What every reader should know about studies using electronic health record data but may be afraid to ask, J. Med. Internet Res. 23 (3) (2021) e22219, http://dx.doi.org/10.2196/22219.

[7] G. Hripcsak, D.J. Albers, Next-generation phenotyping of electronic health records, J. Am. Med. Inform. Assoc. 20 (1) (2013) 117–121, Publisher: BMJ Group.

[8] P. Yadav, M. Steinbach, V. Kumar, G. Simon, Mining electronic health records (EHRs) a survey, ACM Comput. Surv. 50 (6) (2018) 1–40, Publisher: ACM New York, NY, USA.

[9] S. Esteban, M. Rodríguez Tablado, F.E. Peper, Y.S. Mahumud, R.I. Ricci, K.S. Kopitowski, S.A. Terrasa, Development and validation of various phenotyping algorithms for Diabetes Mellitus using data from electronic health records, Comput. Methods Programs Biomed. 152 (2017) 53–70, http://dx.doi.org/10.1016/j.cmpb.2017.09.009, URL https://www.sciencedirect.com/science/article/pii/S0169260717300986.

[10] J.M. Banda, M. Seneviratne, T. Hernandez-Boussard, N.H. Shah, Advances in electronic phenotyping: from rule-based definitions to machine learning models, Annu. Rev. Biomed. Data Sci. 1 (2018) 53–68, Publisher: Annual Reviews.

[11] Y. Bengio, A. Courville, P. Vincent, Representation learning: A review and new perspectives, IEEE Trans. Pattern Anal. Mach. Intell. 35 (8) (2013) 1798–1828, Publisher: IEEE.

[12] R. Miotto, F. Wang, S. Wang, X. Jiang, J.T. Dudley, Deep learning for healthcare: review, opportunities and challenges, Brief. Bioinform. 19 (6) (2018) 1236–1246, http://dx.doi.org/10.1093/bib/bbx044.

[13] W.-H. Weng, P. Szolovits, Representation learning for electronic health records, 2019, arXiv:1909.09248 [Cs, Stat], URL http://arxiv.org/abs/1909.09248.

[14] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A.N. Gomez, L. Kaiser, I. Polosukhin, Attention is all you need, 2017, arXiv preprint arXiv:1706.03762.

[15] Y. Chang, X. Wang, J. Wang, Y. Wu, L. Yang, K. Zhu, H. Chen, X. Yi, C. Wang, Y. Wang, W. Ye, Y. Zhang, Y. Chang, P.S. Yu, Q. Yang, X. Xie, A survey on evaluation of large language models, 15, (3) (ISSN: 2157-6904) 2024, pp. 39:1–39:45, http://dx.doi.org/10.1145/3641289,

[16] D. Van Veen, C. Van Uden, L. Blankemeier, J.-B. Delbrouck, A. Aali, C. Bluethgen, A. Pareek, M. Polacin, E.P. Reis, A. Seehofnerová, N. Rohatgi, P. Hosamani, W. Collins, N. Ahuja, C.P. Langlotz, J. Hom, S. Gatidis, J. Pauly, A.S. Chaudhari, Adapted large language models can outperform medical experts in clinical text summarization, Nature Med. (2024) 1–9, http://dx.doi.org/10.1038/s41591-024-02855-5, URL https://www.nature.com/articles/s41591-024-02855-5, Publisher: Nature Publishing Group.

[17] T. Tu, A. Palepu, M. Schaekermann, K. Saab, J. Freyberg, R. Tanno, A. Wang, B. Li, M. Amin, N. Tomasev, S. Azizi, K. Singhal, Y. Cheng, L. Hou, A. Webson, K. Kulkarni, S.S. Mahdavi, C. Semturs, J. Gottweis, J. Barral, K. Chou, G.S. Corrado, Y. Matias, A. Karthikesalingam, V. Natarajan, Towards conversational diagnostic AI, 2024, http://dx.doi.org/10.48550/arXiv.2401.05654, arXiv, URL http://arxiv.org/abs/2401.05654, arXiv:2401.05654 [cs].

[18] L. Rasmy, Y. Xiang, Z. Xie, C. Tao, D. Zhi, Med-BERT: pretrained contextualized embeddings on large-scale structured electronic health records for disease prediction, Npj Digit. Med. 4 (1) (2021) 1–13, http://dx.doi.org/10.1038/s41746-021-00455-y, URL https://www.nature.com/articles/s41746-021-00455-y, Number: 1 Publisher: Nature Publishing Group.

[19] E. Alsentzer, M.J. Rasmussen, R. Fontoura, A.L. Cull, B. Beaulieu-Jones, K.J. Gray, D.W. Bates, V.P. Kovacheva, Zero-shot interpretable phenotyping of postpartum hemorrhage using large language models, Npj Digit. Med. 6 (1) (2023) 1–10, http://dx.doi.org/10.1038/s41746-023-00957-x, URL https://www.nature.com/articles/s41746-023-00957-x, Number: 1 Publisher: Nature Publishing Group.

[20] C. Yan, H.H. Ong, M.E. Grabowska, M.S. Krantz, W.-C. Su, A.L. Dickson, J.F. Peterson, Q. Feng, D.M. Roden, C.M. Stein, V.E. Kerchberger, B.A. Malin, W.-Q. Wei, Large language models facilitate the generation of electronic health record phenotyping algorithms, J. Am. Med. Inform. Assoc. : JAMIA 31 (9) (2024) 1994–2001, http://dx.doi.org/10.1093/jamia/ocae072, URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC11339509/.

[21] W.E. Thompson, D.M. Vidmar, J.K. De Freitas, J.M. Pfeifer, B.K. Fornwalt, R. Chen, G. Altay, K. Manghnani, A.C. Nelsen, K. Morland, M.C. Stumpe, R. Miotto, Large language models with retrieval-augmented generation for zero-shot disease phenotyping, 2023, http://dx.doi.org/10.48550/arXiv.2312.06457, arXiv, URL http://arxiv.org/abs/2312.06457, arXiv:2312.06457 [cs].

[22] D. Li, A. Kadav, A. Gao, R. Li, R. Bourgon, Automated clinical data extraction with knowledge conditioned LLMs, 2024, http://dx.doi.org/10.48550/arXiv.2406.18027, arXiv, URL http://arxiv.org/abs/2406.18027, arXiv:2406.18027 [cs].

[23] A. Goel, A. Gueta, O. Gilon, C. Liu, S. Erell, L.H. Nguyen, X. Hao, B. Jaber, S. Reddy, R. Kartha, J. Steiner, I. Laish, A. Feder, LLMs accelerate annotation for medical information extraction, 2023, http://dx.doi.org/10.48550/arXiv.2312.02296, arXiv, URL http://arxiv.org/abs/2312.02296, arXiv:2312.02296 [cs].

[24] A. Bayão, A medicina interna do hospital da Luz Lisboa, Med. Interna 28 (1) (2021) 9–12, URL https://revista.spmi.pt/index.php/rpmi/article/download/80/27.

[25] B. Neves, E.D. Haghighi, H.V. Pereira, F. Costa, J.S. Carlos, D. Ferreira, P. Moreno, P.M. Ferreira, J. Machado, B. Goncalves, J.M. Moreira, F. Leite, N.A. da Silva, Impact of a wearable-based physical activity and sleep intervention in multimorbidity patients: protocol for a randomized controlled trial, BMC Geriatr. 23 (1) (2023) 853, http://dx.doi.org/10.1186/s12877-023-04511-y, URL https://bmcgeriatr.biomedcentral.com/articles/10.1186/s12877-023-04511-y.

[26] M. Fortin, J. Almirall, K. Nicholson, Development of a Research Tool to Document Self-Reported Chronic Conditions in Primary Care, SAGE Publications Sage UK, London, England, 2017.

[27] N.L. Pratt, M. Kerr, J.D. Barratt, A. Kemp-Casey, L.M.K. Ellett, E. Ramsay, E.E. Roughead, The validity of the RX-risk comorbidity index using medicines mapped to the Anatomical Therapeutic Chemical (ATC) classification system, BMJ Open 8 (4) (2018) e021122, http://dx.doi.org/10.1136/bmjopen-2017-021122, URL https://bmjopen.bmj.com/content/8/4/e021122, Publisher: British Medical Journal Publishing Group Section: Epidemiology.

[28] J. Wang, E. Shi, S. Yu, Z. Wu, C. Ma, H. Dai, Q. Yang, Y. Kang, J. Wu, H. Hu, C. Yue, H. Zhang, Y. Liu, X. Li, B. Ge, D. Zhu, Y. Yuan, D. Shen, T. Liu, S. Zhang, Prompt engineering for healthcare: Methodologies and applications, 2023, arXiv, URL http://arxiv.org/abs/2304.14670, arXiv:2304.14670 [cs].

[29] S. Sivarajkumar, Y. Wang, HealthPrompt: A zero-shot learning paradigm for clinical natural language processing, AMIA ... Annu. Symp. Proc. AMIA Symp. 2022 (2022) 972–981.

[30] T.B. Brown, B. Mann, N. Ryder, M. Subbiah, J. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, S. Agarwal, A. Herbert-Voss, G. Krueger, T. Henighan, R. Child, A. Ramesh, D.M. Ziegler, J. Wu, C. Winter, C. Hesse, M. Chen, E. Sigler, M. Litwin, S. Gray, B. Chess, J. Clark, C. Berner, S. McCandlish, A. Radford, I. Sutskever, D. Amodei, Language models are few-shot learners, 2020, http://dx.doi.org/10.48550/arXiv.2005.14165, arXiv, URL http://arxiv.org/abs/2005.14165, arXiv:2005.14165 [cs].

[31] OpenAI, J. Achiam, S. Adler, S. Agarwal, L. Ahmad, I. Akkaya, F.L. Aleman, D. Almeida, J. Altenschmidt, S. Altman, S. Anadkat, R. Avila, I. Babuschkin, S. Balaji, V. Balcom, P. Baltescu, H. Bao, M. Bavarian, J. Belgum, I. Bello, J. Berdine, G. Bernadett-Shapiro, C. Berner, L. Bogdonoff, O. Boiko, M. Boyd, A.-L. Brakman, G. Brockman, T. Brooks, M. Brundage, K. Button, T. Cai, R. Campbell, A. Cann, B. Carey, C. Carlson, R. Carmichael, B. Chan, C. Chang, F. Chantzis, D. Chen, S. Chen, R. Chen, J. Chen, M. Chen, B. Chess, C. Cho, C. Chu, H.W. Chung, D. Cummings, J. Currier, Y. Dai, C. Decareaux, T. Degry, N. Deutsch, D. Deville, A. Dhar, D. Dohan, S. Dowling, S. Dunning, A. Ecoffet, A. Eleti, T. Eloundou, D. Farhi, L. Fedus, N. Felix, S.P. Fishman, J. Forte, I. Fulford, L. Gao, E. Georges, C. Gibson, V. Goel, T. Gogineni, G. Goh, R. Gontijo-Lopes, J. Gordon, M. Grafstein, S. Gray, R. Greene, J. Gross, S.S. Gu, Y. Guo, C. Hallacy, J. Han, J. Harris, Y. He, M. Heaton, J. Heidecke, C. Hesse, A. Hickey, W. Hickey, P. Hoeschele, B. Houghton, K. Hsu, S. Hu, X. Hu, J. Huizinga, S. Jain, S. Jain, J. Jang, A. Jiang, R. Jiang, H. Jin, D. Jin, S. Jomoto, B. Jonn, H. Jun, T. Kaftan, L.u. Kaiser, A. Kamali, I. Kanitscheider, N.S. Keskar, T. Khan, L. Kilpatrick, J.W. Kim, C. Kim, Y. Kim, J.H. Kirchner, J. Kiros, M. Knight, D. Kokotajlo, L.u. Kondraciuk, A. Kondrich, A. Konstantinidis, K. Kosic, G. Krueger, V. Kuo, M. Lampe, I. Lan, T. Lee, J. Leike, J. Leung, D. Levy, C.M. Li, R. Lim, M. Lin, S. Lin, M. Litwin, T. Lopez, R. Lowe, P. Lue, A. Makanju, K. Malfacini, S. Manning, T. Markov, Y. Markovski, B. Martin, K. Mayer, A. Mayne, B. McGrew, S.M. McKinney, C. McLeavey, P. McMillan, J. McNeil, D. Medina, A. Mehta, J. Menick, L. Metz, A. Mishchenko, P. Mishkin, V. Monaco, E. Morikawa, D. Mossing, T. Mu, M. Murati, O. Murk, D. Mély, A. Nair, R. Nakano, R. Nayak, A. Neelakantan, R. Ngo, H. Noh, L. Ouyang, C. O'Keefe, J. Pachocki, A. Paino, J. Palermo, A. Pantuliano, G. Parascandolo, J. Parish, E. Parparita, A. Passos, M. Pavlov, A. Peng, A. Perelman, F.d.B. Peres, M. Petrov, H.P.d. Pinto, Michael, Pokorny, M. Pokrass, V.H. Pong, T. Powell, A. Power, B. Power, E. Proehl, R. Puri, A. Radford, J. Rae, A. Ramesh, C. Raymond, F. Real, K. Rimbach, C. Ross, B. Rotsted, H. Roussez, N. Ryder, M. Saltarelli, T. Sanders, S. Santurkar, G. Sastry, H. Schmidt, D. Schnurr, J. Schulman, D. Selsam, K. Sheppard, T. Sherbakov, J. Shieh, S. Shoker, P. Shyam, S. Sidor, E. Sigler, M. Simens, J. Sitkin, K. Slama, I. Sohl, B. Sokolowsky, Y. Song, N. Staudacher, F.P. Such, N. Summers, I. Sutskever, J. Tang, N. Tezak, M.B. Thompson, P. Tillet, A. Tootoonchian, E. Tseng, P. Tuggle, N. Turley, J. Tworek, J.F.C. Uribe, A. Vallone, A. Vijayvergiya, C. Voss, C. Wainwright, J.J. Wang, A. Wang, B. Wang, J. Ward, J. Wei, C.J. Weinmann, A. Welihinda, P. Welinder, J. Weng, L. Weng, M. Wiethoff, D. Willner, C. Winter, S. Wolrich, H. Wong, L. Workman, S. Wu, J. Wu, M. Wu, K. Xiao, T. Xu, S. Yoo, K. Yu, Q. Yuan, W. Zaremba, R. Zellers, C. Zhang, M. Zhang, S. Zhao, T. Zheng, J. Zhuang, W. Zhuk, B. Zoph, GPT-4 technical report, 2024, http://dx.doi.org/10.48550/arXiv.2303.08774, arXiv, URL http://arxiv.org/abs/2303.08774, arXiv:2303.08774 [cs].

[32] A. Dubey, A. Jauhri, A. Pandey, A. Kadian, A. Al-Dahle, A. Letman, A. Mathur, A. Schelten, A. Yang, A. Fan, A. Goyal, A. Hartshorn, A. Yang, A. Mitra, A. Sravankumar, A. Korenev, A. Hinsvark, A. Rao, A. Zhang, A. Rodriguez, A. Gregerson, A. Spataru, B. Roziere, B. Biron, B. Tang, B. Chern, C. Caucheteux, C. Nayak, C. Bi, C. Marra, C. McConnell, C. Keller, C. Touret, C. Wu, C. Wong, C.C. Ferrer, C. Nikolaidis, D. Allonsius, D. Song, D. Pintz, D. Livshits, D. Esiobu, D. Choudhary, D. Mahajan, D. Garcia-Olano, D. Perino, D. Hupkes, E. Lakomkin, E. AlBadawy, E. Lobanova, E. Dinan, E.M. Smith, F. Radenovic, F. Zhang, G. Synnaeve, G. Lee, G.L. Anderson, G. Nail, G. Mialon, G. Pang, G. Cucurell, H. Nguyen, H. Korevaar, H. Xu, H. Touvron, I. Zarov, I.A. Ibarra, I. Kloumann, I. Misra, I. Evtimov, J. Copet, J. Lee, J. Geffert, J. Vranes, J. Park, J. Mahadeokar, J. Shah, J. van der Linde, J. Billock, J. Hong, J. Lee, J. Fu, J. Chi, J. Huang, J. Liu, J. Wang, J. Yu, J. Bitton, J. Spisak, J. Park, J. Rocca, J. Johnstun, J. Saxe, J. Jia, K.V. Alwala, K. Upasani, K. Plawiak, K. Li, K. Heafield, K. Stone, K. El-Arini, K. Iyer, K. Malik, K. Chiu, K. Bhalla, L. Rantala-Yeary, L. van der Maaten, L. Chen,

L. Tan, L. Jenkins, L. Martin, L. Madaan, L. Malo, L. Blecher, L. Landzaat, L. de Oliveira, M. Muzzi, M. Pasupuleti, M. Singh, M. Paluri, M. Kardas, M. Oldham, M. Rita, M. Pavlova, M. Kambadur, M. Lewis, M. Si, M.K. Singh, M. Hassan, N. Goyal, N. Torabi, N. Bashlykov, N. Bogoychev, N. Chatterji, O. Duchenne, O. Çelebi, P. Alrassy, P. Zhang, P. Li, P. Vasic, P. Weng, P. Bhargava, P. Dubal, P. Krishnan, P.S. Koura, P. Xu, Q. He, Q. Dong, R. Srinivasan, R. Ganapathy, R. Calderer, R.S. Cabral, R. Stojnic, R. Raileanu, R. Girdhar, R. Patel, R. Sauvestre, R. Polidoro, R. Sumbaly, R. Taylor, R. Silva, R. Hou, R. Wang, S. Hosseini, S. Chennabasappa, S. Singh, S. Bell, S.S. Kim, S. Edunov, S. Nie, S. Narang, S. Raparthy, S. Shen, S. Wan, S. Bhosale, S. Zhang, S. Vandenhende, S. Batra, S. Whitman, S. Sootla, S. Collot, S. Gururangan, S. Borodinsky, T. Herman, T. Fowler, T. Sheasha, T. Georgiou, T. Scialom, T. Speckbacher, T. Mihaylov, T. Xiao, U. Karn, V. Goswami, V. Gupta, V. Ramanathan, V. Kerkez, V. Gonguet, V. Do, V. Vogeti, V. Petrovic, W. Chu, W. Xiong, W. Fu, W. Meers, X. Martinet, X. Wang, X.E. Tan, X. Xie, X. Jia, X. Wang, Y. Goldschlag, Y. Gaur, Y. Babaei, Y. Wen, Y. Song, Y. Zhang, Y. Li, Y. Mao, Z.D. Coudert, Z. Yan, Z. Chen, Z. Papakipos, A. Singh, A. Grattafiori, A. Jain, A. Kelsey, A. Shajnfeld, A. Gangidi, A. Victoria, A. Goldstand, A. Menon, A. Sharma, A. Boesenberg, A. Vaughan, A. Baevski, A. Feinstein, A. Kallet, A. Sangani, A. Yunus, A. Lupu, A. Alvarado, A. Caples, A. Gu, A. Ho, A. Poulton, A. Ryan, A. Ramchandani, A. Franco, A. Saraf, A. Chowdhury, A. Gabriel, A. Bharambe, A. Eisenman, A. Yazdan, B. James, B. Maurer, B. Leonhardi, B. Huang, B. Loyd, B. De Paola, B. Paranjape, B. Liu, B. Wu, B. Ni, B. Hancock, B. Wasti, B. Spence, B. Stojkovic, B. Gamido, B. Montalvo, C. Parker, C. Burton, C. Mejia, C. Wang, C. Kim, C. Zhou, C. Hu, C.-H. Chu, C. Cai, C. Tindal, C. Feichtenhofer, D. Civin, D. Beaty, D. Kreymer, D. Li, D. Wyatt, D. Adkins, D. Xu, D. Testuggine, D. David, D. Parikh, D. Liskovich, D. Foss, D. Wang, D. Le, D. Holland, E. Dowling, E. Jamil, E. Montgomery, E. Presani, E. Hahn, E. Wood, E. Brinkman, E. Arcaute, E. Dunbar, E. Smothers, F. Sun, F. Kreuk, F. Tian, F. Ozgenel, F. Caggioni, F. Guzmán, F. Kanayet, F. Seide, G.M. Florez, G. Schwarz, G. Badeer, G. Swee, G. Halpern, G. Thattai, G. Herman, G. Sizov, Guangyi, Zhang, G. Lakshminarayanan, H. Shojanazeri, H. Zou, H. Wang, H. Zha, H. Habeeb, H. Rudolph, H. Suk, H. Aspegren, H. Goldman, I. Damlaj, I. Molybog, I. Tufanov, I.-E. Veliche, I. Gat, J. Weissman, J. Geboski, J. Kohli, J. Asher, J.-B. Gaya, J. Marcus, J. Tang, J. Chan, J. Zhen, J. Reizenstein, J. Teboul, J. Zhong, J. Jin, J. Yang, J. Cummings, J. Carvill, J. Shepard, J. McPhie, J. Torres, J. Ginsburg, J. Wang, K. Wu, K.H. U, K. Saxena, K. Prasad, K. Khandelwal, K. Zand, K. Matosich, K. Veeraraghavan, K. Michelena, K. Li, K. Huang, K. Chawla, K. Lakhotia, K. Huang, L. Chen, L. Garg, L. A, L. Silva, L. Bell, L. Zhang, L. Guo, L. Yu, L. Moshkovich, L. Wehrstedt, M. Khabsa, M. Avalani, M. Bhatt, M. Tsimpoukelli, M. Mankus, M. Hasson, M. Lennie, M. Reso, M. Groshev, M. Naumov, M. Lathi, M. Keneally, M.L. Seltzer, M. Valko, M. Restrepo, M. Patel, M. Vyatskov, M. Samvelyan, M. Clark, M. Macey, M. Wang, M.J. Hermoso, M. Metanat, M. Rastegari, M. Bansal, N. Santhanam, N. Parks, N. White, N. Bawa, N. Singhal, N. Egebo, N. Usunier, N.P. Laptev, N. Dong, N. Zhang, N. Cheng, O. Chernoguz, O. Hart, O. Salpekar, O. Kalinli, P. Kent, P. Parekh, P. Saab, P. Balaji, P. Rittner, P. Bontrager, P. Roux, P. Dollar, P. Zvyagina, P. Ratanchandani, P. Yuvraj, Q. Liang, R. Alao, R. Rodriguez, R. Ayub, R. Murthy, R. Nayani, R. Mitra, R. Li, R. Hogan, R. Battey, R. Wang, R. Maheswari, R. Howes, R. Rinott, S.J. Bondu, S. Datta, S. Chugh, S. Hunt, S. Dhillon, S. Sidorov, S. Pan, S. Verma, S. Yamamoto, S. Ramaswamy, S. Lindsay, S. Lindsay, S. Feng, S. Lin, S.C. Zha, S. Shankar, S. Zhang, S. Zhang, S. Wang, S. Agarwal, S. Sajuyigbe, S. Chintala, S. Max, S. Chen, S. Kehoe, S. Satterfield, S. Govindaprasad, S. Gupta, S. Cho, S. Virk, S. Subramanian, S. Choudhury, S. Goldman, T. Remez, T. Glaser, T. Best, T. Kohler, T. Robinson, T. Li, T. Zhang, T. Matthews, T. Chou, T. Shaked, V. Vontimitta, V. Ajayi, V. Montanez, V. Mohan, V.S. Kumar, V. Mangla, V. Albiero, V. Ionescu, V. Poenaru, V.T. Mihailescu, V. Ivanov, W. Li, W. Wang, W. Jiang, W. Bouaziz, W. Constable, X. Tang, X. Wang, X. Wu, X. Wang, X. Xia, X. Wu, X. Gao, Y. Chen, Y. Hu, Y. Jia, Y. Qi, Y. Li, Y. Zhang, Y. Zhang, Y. Adi, Y. Nam, Yu, Wang, Y. Hao, Y. Qian, Y. He, Z. Rait, Z. DeVito, Z. Rosnbrick, Z. Wen, Z. Yang, Z. Zhao, The Llama 3 herd of models, 2024, http://dx.doi.org/10.48550/arXiv.2407.21783, arXiv, URL http://arxiv.org/abs/2407.21783, arXiv:2407.21783 [cs].

[33] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, Scikit-learn: Machine learning in Python journal of machine learning research, J. Mach. Learn. Res. 12 (2011) 2825–2830.

[34] J. Zhang, K. Sun, A. Jagadeesh, P. Falakaflaki, E. Kayayan, G. Tao, M. Haghighat Ghahfarokhi, D. Gupta, A. Gupta, V. Gupta, Y. Guo, The potential and pitfalls of using a large language model such as ChatGPT, GPT-4, or LLaMA as a clinical assistant, J. Am. Med. Informatics Assoc.: JAMIA 31 (9) (2024) 1884–1891, http://dx.doi.org/10.1093/jamia/ocae184.

[35] R. Fornasiere, N. Brunello, V. Scotti, M. Carman, Medical information extraction with large language models, in: M. Abbas, A.A. Freihat (Eds.), Proceedings of the 7th International Conference on Natural Language and Speech Processing, ICNLSP 2024, Association for Computational Linguistics, Trento, 2024, pp. 456–466, URL https://aclanthology.org/2024.icnlsp-1.47/.

[36] M. Griot, C. Hemptinne, J. Vanderdonckt, D. Yuksel, Large language models lack essential metacognition for reliable medical reasoning, Nat. Commun. 16 (1) (2025) 642, http://dx.doi.org/10.1038/s41467-024-55628-6, URL https://www.nature.com/articles/s41467-024-55628-6, Publisher: Nature Publishing Group.

[37] M. Steyvers, H. Tejeda, A. Kumar, C. Belem, S. Karny, X. Hu, L.W. Mayer, P. Smyth, What large language models know and what people think they know, Nat. Mach. Intell. (2025) 1–11, http://dx.doi.org/10.1038/s42256-024-00976-7, URL https://www.nature.com/articles/s42256-024-00976-7, Publisher: Nature Publishing Group.

[38] T. Zack, E. Lehman, M. Suzgun, J.A. Rodriguez, L.A. Celi, J. Gichoya, D. Jurafsky, P. Szolovits, D.W. Bates, R.-E.E. Abdulnour, A.J. Butte, E. Alsentzer, Assessing the potential of GPT-4 to perpetuate racial and gender biases in health care: a model evaluation study, Lancet Digit. Heal. 6 (1) (2024) e12–e22, http://dx.doi.org/10.1016/S2589-7500(23)00225-X, URL https://www.thelancet.com/journals/landig/article/PIIS2589-7500(23)00225-X/fulltext, Publisher: Elsevier.

[39] K. Singhal, T. Tu, J. Gottweis, R. Sayres, E. Wulczyn, L. Hou, K. Clark, S. Pfohl, H. Cole-Lewis, D. Neal, M. Schaekermann, A. Wang, M. Amin, S. Lachgar, P. Mansfield, S. Prakash, B. Green, E. Dominowska, B.A.y. Arcas, N. Tomasev, Y. Liu, R. Wong, C. Semturs, S.S. Mahdavi, J. Barral, D. Webster, G.S. Corrado, Y. Matias, S. Azizi, A. Karthikesalingam, V. Natarajan, Towards expert-level medical question answering with large language models, 2023, http://dx.doi.org/10.48550/arXiv.2305.09617, arXiv, URL http://arxiv.org/abs/2305.09617, arXiv:2305.09617 [cs].

[40] L. Wang, A. Wen, S. Fu, X. Ruan, M. Huang, R. Li, Q. Lu, A.E. Williams, H. Liu, Adoption of the OMOP CDM for cancer research using real-world data: Current status and opportunities, 2024, http://dx.doi.org/10.1101/2024.08.23.24311950, 2024.08.23.24311950, MedRxiv, URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC11370549/.

[41] M. Fruchart, P. Quindroit, C. Jacquemont, J.-B. Beuscart, M. Calafiore, A. Lamer, Transforming primary care data into the observational medical outcomes partnership common data model: Development and usability study, JMIR Med. Inform. 12 (1) (2024) e49542, http://dx.doi.org/10.2196/49542, URL https://medinform.jmir.org/2024/1/e49542, Company: JMIR Medical Informatics Distributor: JMIR Medical Informatics Institution: JMIR Medical Informatics Label: JMIR Medical Informatics Publisher: JMIR Publications Inc., Toronto, Canada.