medRxiv preprint doi: https://doi.org/10.1101/2023.12.19.23300230; this version posted December 19, 2023. The copyright holder for this preprint **(which was not certified by peer review)** is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.

It is made available under a CC-BY-NC 4.0 International license .

Title: Large Language Models Facilitate the Generation of Electronic Health Record Phenotyping Algorithms

Authors:

Chao Yan[1], PhD, Henry H. Ong[1], PhD, Monika E. Grabowska[1], MS, Matthew S. Krantz[1], MD, Wu-Chen Su[1], MS, Alyson L. Dickson[2], MA, Josh F. Peterson[1,2], MD, MPH, QiPing Feng[2], PhD, Dan M. Roden[1], MD, C. Michael Stein[2], MD, V. Eric Kerchberger[1,2], MD, Brad A. Malin[1,3,4], PhD, Wei-Qi Wei[1,3], MD, PhD

Affiliations:

[1]Department of Biomedical Informatics, Vanderbilt University Medical Center, Nashville, TN
[2]Department of Medicine, Vanderbilt University Medical Center, Nashville, TN
[3]Department of Computer Science, Vanderbilt University, Nashville, TN
[4]Department of Biostatistics, Vanderbilt University Medical Center, Nashville, TN

Corresponding author:

Wei-Qi Wei, MD, PhD
Department of Biomedical Informatics & Computer Science
Vanderbilt University Medical Center, Vanderbilt University
Suite 1500, 2525 West End Ave, Nashville, TN, 37203
Email: wei-qi.wei@vumc.org

Keywords: Phenotyping, Electronic health records, Large language models, ChatGPT

Word count: 3504

NOTE: This preprint reports new research that has not been certified by peer review and should not be used to guide clinical practice.

medRxiv preprint doi: https://doi.org/10.1101/2023.12.19.23300230; this version posted December 19, 2023. The copyright holder for this preprint **(which was not certified by peer review)** is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.

It is made available under a CC-BY-NC 4.0 International license .

ABSTRACT

Objectives: Phenotyping is a core task in observational health research utilizing electronic health records (EHRs). Developing an accurate algorithm typically demands substantial input from domain experts, involving extensive literature review and evidence synthesis. This burdensome process limits scalability and delays knowledge discovery. We investigate the potential for leveraging large language models (LLMs) to enhance the efficiency of EHR phenotyping by generating drafts of high-quality algorithms.

Materials and Methods: We prompted four LLMs—ChatGPT-4, ChatGPT-3.5, Claude 2, and Bard—in October 2023, asking them to generate executable phenotyping algorithms in the form of SQL queries adhering to a common data model for three clinical phenotypes (i.e., type 2 diabetes mellitus, dementia, and hypothyroidism). Three phenotyping experts evaluated the returned algorithms across several critical metrics. We further implemented the top-rated algorithms from each LLM and compared them against clinician-validated phenotyping algorithms from the Electronic Medical Records and Genomics (eMERGE) network.

Results: ChatGPT-4 and ChatGPT-3.5 exhibited significantly higher overall expert evaluation scores in instruction following, algorithmic logic, and SQL executability, when compared to Claude 2 and Bard. Although ChatGPT-4 and ChatGPT-3.5 effectively identified relevant clinical concepts, they exhibited immature capability in organizing phenotyping criteria with the proper logic, leading to phenotyping algorithms that were either excessively restrictive (with low recall) or overly broad (with low positive predictive values).

Conclusion: Both ChatGPT versions 3.5 and 4 demonstrate the capability to enhance EHR phenotyping efficiency by drafting algorithms of reasonable quality. However, the optimal performance of these algorithms necessitates the involvement of domain experts.

medRxiv preprint doi: https://doi.org/10.1101/2023.12.19.23300230; this version posted December 19, 2023. The copyright holder for this preprint **(which was not certified by peer review)** is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.

It is made available under a CC-BY-NC 4.0 International license .

INTRODUCTION

Electronic health record (EHR) phenotyping, which involves creating algorithms to identify and correctly classify a patient's observable characteristics by integrating complex clinical data, has become pivotal in observational health research[1]. Developing EHR phenotypes is an intricate and labor-intensive process that demands extensive expertise in both the clinical and informatics domains[2,3]. While phenotyping includes the identification of individuals with specific characteristics, it also necessitates the selection of suitable controls for meaningful comparisons with the identified cases[4].

Rule-based computable phenotyping algorithms rely on clinical experts to select specialized criteria (e.g., diagnosis codes, medications, and laboratory values) likely to define a phenotype of interest. These algorithms are tested and tailored for more precise phenotype identification, resulting in better performance compared to general high-throughput methods[5-8]. However, the iterative nature of this process often requires substantial literature review and discussions with clinical experts to generate a single phenotyping algorithm, thereby limiting the scalability of this approach in practice[2]. Furthermore, implementation of phenotyping algorithms by secondary sites requires additional informatics expertise, manual effort, and time to adapt the existing code to different databases and EHR systems.

Recently, large language models (LLMs) have demonstrated effectiveness in information extraction and summarization[9], indicating a potential benefit in phenotyping by reducing the time required for literature review and synthesis during the phenotype generation process. In this preliminary report, we investigate the novel application of LLMs for generating computable phenotyping algorithms to assess whether such tools can effectively expedite EHR phenotype development. We appraised four LLMs—ChatGPT-4[10], ChatGPT-3.5[11], Claude 2[12], and PaLM 2-powered Bard[13]—to generate phenotyping algorithms for three clinical phenotypes—type 2 diabetes mellitus (T2DM), dementia, and hypothyroidism. We subsequently implemented the top-rated algorithms as adjudicated by phenotyping experts using multiple critical metrics from each LLM and compared them against the clinician-validated phenotyping algorithms from the Electronic Medical Records and Genomics (eMERGE) network[14,15].

medRxiv preprint doi: https://doi.org/10.1101/2023.12.19.23300230; this version posted December 19, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.

It is made available under a CC-BY-NC 4.0 International license .

METHODS

We selected four LLMs in their default configurations to test their phenotyping algorithm generation capacity for three common clinical phenotypes. These LLMs were: ChatGPT-4 (by OpenAI)[10], ChatGPT-3.5 (by OpenAI)[11], Claude 2 (by Anthropic)[12], and Bard (by Google, powered by PaLM 2)[13]. These models were chosen because of their widespread use, easy accessibility, extensive evaluation, robust computational capabilities, and proficiency in handling and generating lengthy texts—qualities crucial for sustainably supporting phenotyping tasks[16].

This pilot study specifically focused on three clinical phenotypes: T2DM[17,18], dementia[19], and hypothyroidism[20,21]. We chose these phenotypes because they have existing algorithms that have undergone extensive validation processes and demonstrated highly accurate and consistent performances with well-documented results. Data collection via web-based interactions with the LLMs occurred in October 2023, with subsequent data analysis completed in November 2023. This study was approved by the institutional review boards at Vanderbilt University Medical Center (IRB #: 201434).

Figure 1 illustrates an overview of the study pipeline, which was comprised of two main components—prompting (steps 1-3, highlighted in pink) and evaluating (steps 4-9, highlighted in blue) LLMs.

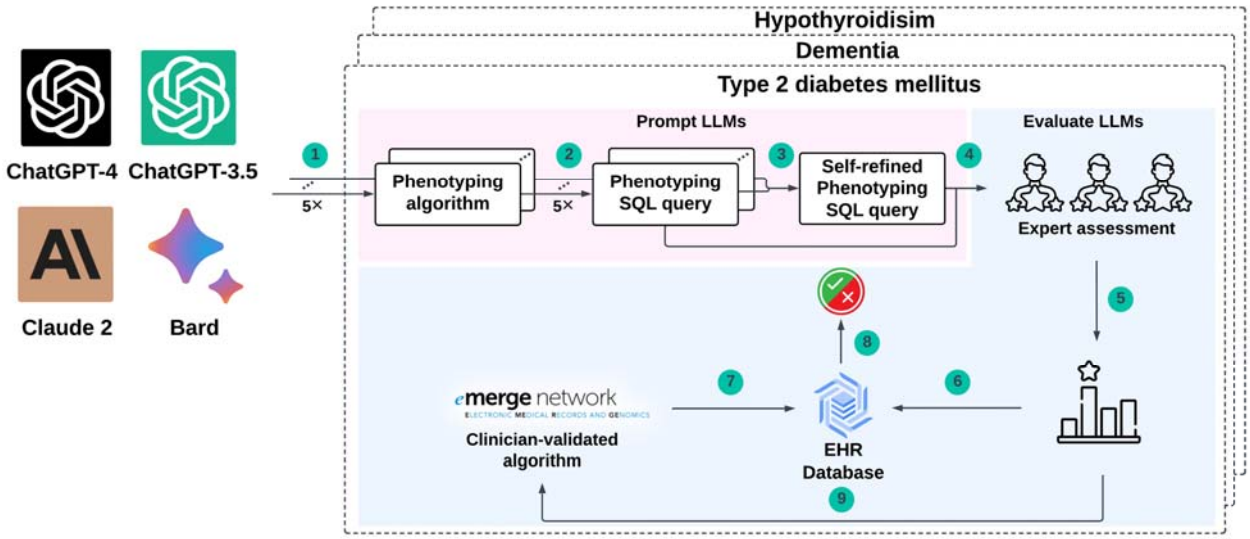


Figure 1. An architectural overview of the study pipeline.

Prompting large language models to generate phenotyping algorithms

medRxiv preprint doi: https://doi.org/10.1101/2023.12.19.23300230; this version posted December 19, 2023. The copyright holder for this preprint **(which was not certified by peer review)** is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.

It is made available under a CC-BY-NC 4.0 International license .

We prompted the LLMs to generate executable SQL queries to identify phenotype cases from structured EHR data organized according to the Observational Medical Outcomes Partnership (OMOP) Common Data Model (CDM), a standard EHR data framework that enables efficient data analysis and sharing across institutions. To achieve this, we designed two distinct prompting strategies, hereafter termed α-prompting and β-prompting.

The α-prompting strategy (steps 1 and 2 in **Figure 1**) had two steps. The first step focused on obtaining a pseudocode version of the phenotyping algorithm (referred to as a pseudo-phenotyping algorithm), which emphasized identifying and integrating critical phenotyping criteria, as well as determining the strategy for combining these criteria. We utilized the chain-of-thought (CoT) prompting strategy[24], an effective method for directing LLMs through a series of reasoning steps to resolve complex problems like humans. Specifically, we framed our instruction to guide reasoning as follows: "Let's think step by step: 1. List the critical criteria to consider. 2. Determine how these criteria should be combined. 3. Derive the final algorithm" (Supplemental **Table 1**). Additionally, multiple detailed instructions were specified so that the produced pseudo-phenotyping algorithm adhered to the OMOP concepts, including diagnosis codes (in ICD-9-CM and ICD-10-CM), symptoms, procedures, laboratory tests, and medications (both generic and brand names). We also mandated that the pseudo-phenotyping algorithm maintain a style consistent with the SQL logic, to facilitate generation of the SQL query in the second step.

Using the response of an LLM in the initial step, the second step involved converting the pseudo-phenotyping algorithm into an executable SQL query (referred to as an SQL-formatted phenotyping algorithm) for implementation in an OMOP CDM-based EHR database for subsequent validation (Supplemental **Table 1**). Due to the probabilistic nature of LLMs, variations in the generated phenotyping algorithms are guaranteed. Consequently, we executed α-prompting five times independently for each phenotype and for each LLM to account for response variability.

The β-prompting strategy (step 3 in **Figure 1**) was designed to first present the LLM with the five SQL-formatted phenotyping algorithms generated from the α-prompting strategy and then instruct the LLM to assess the quality of these algorithms and generate an improved one (Supplemental **Table 1**). This strategy, proven to effectively mitigate hallucinations[25-27],

medRxiv preprint doi: https://doi.org/10.1101/2023.12.19.23300230; this version posted December 19, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.

It is made available under a CC-BY-NC 4.0 International license .

leverages an LLM's ability to evaluate scientific texts based on the extensive knowledge encoded during model pre-training. As a result, the LLM can produce an updated version of the phenotyping algorithm through analytical reflections. We also deployed the CoT strategy in β-prompting, which involved initially identifying the correct, incorrect, and missing criteria for each previously generated algorithm. The β-prompting strategy was executed in an independent session of the LLM (distinct from and subsequent to those used for α-prompting) and was executed only once.

Evaluating the quality of LLM-generated phenotyping algorithms

We then performed a comprehensive analysis, encompassing both qualitative and quantitative assessments, to evaluate the efficacy of the phenotyping algorithms generated by the four different LLMs on the three diseases of focus as described above. For the qualitative analysis, three experts in EHR phenotyping and clinical medicine (W.Q.W., M.E.G., and V.E.K.) independently reviewed and rated the LLM-generated SQL-formatted phenotyping algorithms in a blind manner. We further compared the concepts utilized in these algorithms with those found in clinician-validated phenotyping algorithms developed by the eMERGE network[14,15]. For quantitative assessment, we implemented the top-rated SQL-formatted phenotyping algorithms using EHR data from Vanderbilt University Medical Center (VUMC) and assessed their performance against established eMERGE algorithms[17-21].

Expert assessment Each of the three experts conducted independent reviews for every SQL-formatted phenotyping algorithm (4 LLMs, 3 phenotypes, and 2 strategies, for 24 algorithms in total). The evaluation focused on three dimensions: 1) adherence to instructions, which assessed how well the LLM conformed to predefined formatting rules; 2) the generation of proficient phenotyping algorithms based on knowledge, which evaluated the LLM's ability to synthesize and organize phenotyping-related information effectively; and 3) presentation in executable SQL format, which measured the potential of an LLM to reduce the labor-intensive human efforts required for EHR implementation and validation. Detailed guidelines for expert evaluation can be found in Supplemental **Table 2**. Experts assigned categorical scores ("Good [3]," "Medium [2]," or "Poor [1]") for each axis based on predefined criteria, providing justifications accordingly. Interrater reliability was assessed using the weighted Cohen's Kappa score[28]. We compared LLMs' rated scores using the Wilcoxon signed-rank test[29] with a significance level of 0.05.

medRxiv preprint doi: https://doi.org/10.1101/2023.12.19.23300230; this version posted December 19, 2023. The copyright holder for this preprint **(which was not certified by peer review)** is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.

It is made available under a CC-BY-NC 4.0 International license .

Comparison of concept coverage with eMERGE phenotyping algorithms This analysis provided a comprehensive comparison of the concepts within phenotyping algorithms generated by LLMs and established EHR algorithms. We systematically reviewed and compared all the concepts employed in the algorithms for diagnoses, laboratory tests, procedures, medications, symptoms, and exclusions, and then summarized the noteworthy findings.

Implementation of LLM-generated algorithms and performance evaluation We deployed the highest-rated phenotyping algorithms for each phenotype in a research cohort at VUMC (n=84,821). This cohort, extensively utilized in phenotyping research, has been a significant resource for phenotypic studies[30,31]. As a benchmark, we implemented three eMERGE algorithms updated with current ICD-10-CM codes[31] to identify phenotype cases and controls. The cases and controls identified by the eMERGE algorithms served as a reference standard to assess the effectiveness of the LLM-generated algorithms. In our subsequent analysis, we excluded patients not categorized as either case or control by the eMERGE algorithm, as their data did not meet the criteria for either situation. Each of the top-rated LLM-generated algorithms required some modifications to be executable in a cloud-based platform to securely query VUMC's research clinical databases that follow the OMOP CDM. We limited changes to technical domain knowledge, as opposed to clinical domain knowledge. For example, we edited the database names, but did not edit drug names or ICD codes.

We used the following metrics for evaluation: 1) positive predictive values (PPV), defined as the number of cases mutually identified by the eMERGE algorithm and an LLM over the total number of identified cases by the LLM; 2) recall, defined as the number of cases mutually identified by the eMERGE algorithm and an LLM over the total number of cases identified by the eMERGE algorithm; and 3) false positive rate (FPR), defined as the number of patients identified as cases by an LLM but identified as controls by the eMERGE algorithm (false positives) over the sum of false positives and number of cases identified by the eMERGE algorithm.

medRxiv preprint doi: https://doi.org/10.1101/2023.12.19.23300230; this version posted December 19, 2023. The copyright holder for this preprint **(which was not certified by peer review)** is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.

It is made available under a CC-BY-NC 4.0 International license .

RESULTS

Expert assessments

The average interrater reliability was 0.59 [0.50-0.68], indicating a moderate to substantial agreement among experts, considering the categorical nature of the data (rather than dichotomous) and the variations in scoring criteria among different experts[32].

By mapping the experts' assessments of "Good", "Medium", or "Poor" to numerical scores of 3, 2, and 1, respectively, ChatGPT-4 (mean [95% confidence interval]: 2.57 [2.40-2.75]) and ChatGPT-3.5 (2.43 [2.25-2.60]) exhibited significantly higher overall expert evaluation scores than Claude 2 (1.91 [1.68-2.13]) and Bard (1.20 [1.09-1.31]) (**Figure 2A**). ChatGPT-4 marginally outperformed ChatGPT-3.5, though the differences were not statistically significant. Moreover, the β-prompting strategy did not significantly differ from the α-prompting, according to experts' evaluation (**Figure 2B**). Furthermore, experts assigned higher scores to LLMs for their effectiveness in generating phenotyping algorithms for T2DM and hypothyroidism compared to dementia (**Figure 2C**). The radar plot shown in **Figure 2D** displays the average scores for each involved LLM across the three axes of evaluation. There are two key findings. First, ChatGPT-4 and ChatGPT-3.5 were rated consistently better than Claude 2 and Bard in following instructions, algorithmic logic, and SQL executability. Second, ChatGPT-4 was considered to be on par with ChatGPT-3.5 in its ability to follow instructions and SQL executability, yet it surpassed ChatGPT-3.5 in its algorithmic logic. Considering these findings, the remainder of our results focused exclusively on ChatGPT-4 and ChatGPT-3.5.

medRxiv preprint doi: https://doi.org/10.1101/2023.12.19.23300230; this version posted December 19, 2023. The copyright holder for this preprint **(which was not certified by peer review)** is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.

It is made available under a CC-BY-NC 4.0 International license .

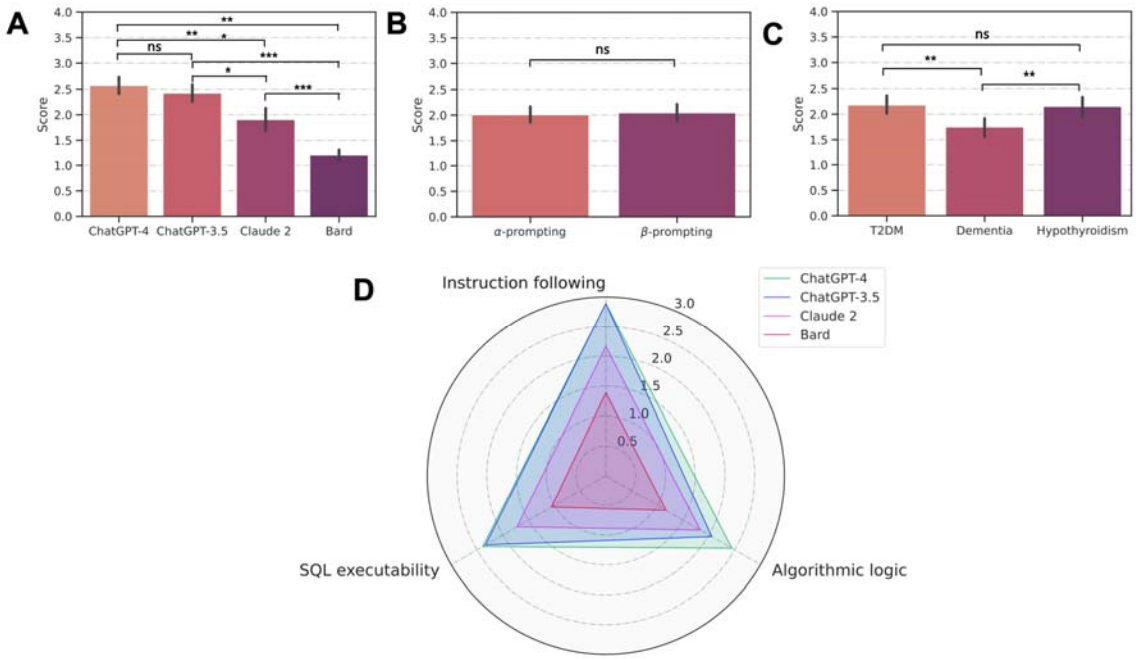


Figure 2. A comparative analysis based on expert evaluations focusing on A) four large language models, B) two prompting strategies, C) three phenotypes, and D) three individual evaluation axes. Numeric scores of 3, 2, and 1 correspond to expert assessments of "Good", "Medium", and "Poor", respectively. ***, **, and * denote *p*<0.001, *p*<0.01, and *p*<0.05, respectively. ns=not significant.

Comparison with eMERGE phenotyping algorithms

Components Table 1 summarizes the clinical concepts identified by the eMERGE phenotyping algorithms and LLM-produced algorithms. A full comparison of concepts can be found in Supplemental **Table** 3. Given that the phenotyping algorithms produced by both ChatGPT-4 and ChatGPT-3.5 from the β-prompting strategy were rated similarly to those from the α-prompting strategy, we focused further analyses on the β-prompting results.

There are several notable observations. For the T2DM phenotyping algorithm, ChatGPT-4 and ChatGPT-3.5 identified relevant diagnosis codes (both ICD-9-CM and ICD-10-CM), lab tests (hemoglobin A1c, fasting blood glucose), and two generic medications (metformin and glipizide) that were also used in the eMERGE phenotyping algorithm. The eMERGE phenotyping algorithm also included medications not identified by either ChatGPT-4 or ChatGPT-3.5 (n=27). Only ChatGPT-4 and ChatGPT-3.5 included symptoms; both eMERGE and ChatGPT-4 provided exclusionary criteria (ICD codes). Notably, the ChatGPT-3.5 model incorrectly included ICD-9-CM codes for type 1 diabetes and applied an incorrect cutoff value for fasting blood glucose (>6.5 mg/dL instead of >125 mg/dL).

medRxiv preprint doi: https://doi.org/10.1101/2023.12.19.23300230; this version posted December 19, 2023. The copyright holder for this preprint **(which was not certified by peer review)** is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.

It is made available under a CC-BY-NC 4.0 International license .

For the dementia phenotyping algorithm, ChatGPT-4 and ChatGPT-3.5 included relevant diagnosis codes (ICD-9-CM and ICD-10-CM, as well as one ICD-10). While there were several overlapping diagnosis codes between each pair of phenotyping algorithms, no diagnosis codes were shared across all three algorithms. ChatGPT-4 and ChatGPT-3.5 included symptoms potentially related to dementia while the eMERGE algorithm did not. All three phenotyping algorithms shared four medications (two generic: donepezil and memantine, and two brand names: Aricept and Namenda), although the eMERGE algorithm specified additional medications that did not appear in either LLM-generated algorithm (n=7). Only the ChatGPT-4 phenotyping algorithm included exclusion criteria (e.g., vitamin B12 level).

For the hypothyroidism phenotyping algorithm, ChatGPT-4 and ChatGPT-3.5 identified relevant diagnosis codes (both ICD-9-CM and ICD-10-CM), lab tests (thyroid stimulating hormone), and medications (generic levothyroxine and brand name Synthroid) used in the eMERGE algorithm. Only the eMERGE algorithm used thyroid autoantibodies (e.g., thyroid antiperoxidase). Notably, the ChatGPT-4 algorithm specified the largest number of medications (n=15), including 9 medications that were not included in either the eMERGE algorithm or the ChatGPT-3.5 algorithm. Once again, only ChatGPT-4 and ChatGPT-3.5 included symptoms. The eMERGE algorithm included 113 exclusions, including 18 ICD codes, 79 CPT codes, and 16 medications, while ChatGPT-4 and ChatGPT-3.5 specified only 2 and 0 exclusions, respectively.

Table 1. Shared and unshared concepts from the eMERGE, ChatGPT-4, and ChatGPT-3.5 phenotyping algorithms for T2DM, dementia, and hypothyroidism. Darker blue shading represents a relatively higher count of unshared concepts when comparing the algorithms. Orange shading represents inaccurate concepts or laboratory cutoffs.

Type 2 diabetes mellitus				
Concept	**Shared**	**Unshared concept count (example)**		
Algorithm	**All**	**eMERGE**	**ChatGPT-4 (β)**	**ChatGPT-3.5 (β)**
Diagnoses	250.*0, 250.*2, E11.*	2 (O24.11)	0	31[a] (250.01)
Lab tests or procedures	HgbA1c[b], Fasting BG[c]	1 Lab (Random BG >200)[d]	1 Lab (Oral GTT >200)[d]	1 Lab (BMI ≥25)[d]
Medications	Metformin, Glipizide	34 (Exenatide)[d]	7 (Glucophage)	1 (Sitagliptin)
Symptoms	None	0	3 (Polyuria)	3 (Polydipsia)
Exclusion by Type	None	9 ICD (250.*1)	3 ICD (250.*3)	0
Dementia				
Concept	**Shared**	**Unshared concept count**		

medRxiv preprint doi: https://doi.org/10.1101/2023.12.19.23300230; this version posted December 19, 2023. The copyright holder for this preprint **(which was not certified by peer review)** is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.

It is made available under a CC-BY-NC 4.0 International license .

		(example)		
Algorithm	**All**	**eMERGE**	**ChatGPT-4 (β)**	**ChatGPT-3.5 (β)**
Diagnoses	None	45 (290.0)[d]	6	20
Lab tests or procedures	None	0	4 Procedures (MMSE)[d]	3 Procedures (MRI scan)[d]
Medications	Donepezil, Aricept, Memantine, Namenda	11 (Cognex)[d]	4 (Galantamine)	0
Symptoms	None	0	3 (Impaired reasoning)[d]	5 (Cognitive impairment)[d]
Exclusion by Type	None	0	3 Labs (B12 level)[d]	0
Hypothyroidism				
Concept	**Shared**	**Unshared concept count (example)**		
Algorithm	**All**	**eMERGE**	**ChatGPT-4 (β)**	**ChatGPT-3.5 (β)**
Diagnoses	244.9, E03.8, E03.9	18 (244)[d]	7 (244.1)[d]	6 (E03.5)[d]
Lab tests or procedures	TSH[e]	4 Labs (Anti-TPO)[d]	1 Lab (Serum free thyroxine)	0
Medications	Levothyroxine, Synthroid	9 (Liothyronine)[d]	15 (Amiodarone)[d,f]	0
Symptoms	None	0	8 (Hair loss)[d]	8 (Constipation)[d]
Exclusion by Type	None	18 ICD (193*)[d] 79 CPT (60240)[d] 16 Medications (Amiodarone)[d]	2 Procedures[g] (Thyroid surgery)	0

Abbreviations: BG, blood glucose; BMI, body mass index; CPT, Current Procedural Terminology; GTT, glucose tolerance test; HgbA1c, hemoglobin A1c; ICD, International Classification of Diseases; MMSE, Mini-Mental State Exam; MRI, magnetic resonance imaging; TPO, thyroperoxidase.

Implementation and evaluation in VUMC Along with the eMERGE algorithms, we successfully deployed all algorithms generated by ChatGPT-4 and ChatGPT-3.5 from the β-prompting strategy except the T2DM algorithm generated by ChatGPT-3.5 (**Table 2**). The failure of this phenotyping algorithm stemmed from its restrictive logic which accumulated LEFT JOINs across various tables, including the Person, Condition_Occurrence, Measurement, Drug_Exposure, and Observation tables. The algorithm required a LEFT JOIN on a list of people who had a record of symptoms in the Observation table (Polyuria, Polydipsia, Unexplained weight loss) and had a value of zero in the "value_as_concept_id" column. A value of zero did not exist for these symptoms, therefore the algorithm could not identify any individuals who met this criterion for T2DM.

The case (and control) prevalences for the eMERGE algorithms in our population were 10.9% (28.0%), 3.5% (91.5%), and 2.4% (30.4%) for T2DM, dementia, and hypothyroidism,

medRxiv preprint doi: https://doi.org/10.1101/2023.12.19.23300230; this version posted December 19, 2023. The copyright holder for this preprint **(which was not certified by peer review)** is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.

It is made available under a CC-BY-NC 4.0 International license .

respectively. The ChatGPT-4 generated dementia algorithm achieved the highest PPV of 96.3% at the expense of a low recall (24.4%). This result was primarily attributable to the relatively restrictive inclusion criteria, which required medication prescription coupled with the co-occurrence of either diagnosis codes, symptoms, or cognitive assessment tests or procedures. The ChatGPT-4 generated algorithms for T2DM and hypothyroidism, as well as the ChatGPT-3.5 generated algorithm for hypothyroidism, achieved high recall but at the cost of lower PPV. In contrast, the ChatGPT-3.5 generated algorithm for dementia achieved balanced PPV and recall.

Table 2. Performance of the phenotyping algorithms generated by ChatGPT-4 and ChatGPT-3.5 from the β-prompting strategy when applied to VUMC data, as measured against clinician-validated algorithms for the eMERGE phenotype cases and controls.

Disease	eMERGE		ChatGPT-4					ChatGPT-3.5				
	True cases	True controls	TP	FP	PPV	Recall	FPR	TP	FP	PPV	Recall	FPR
T2DM	9,293	23,754	8,978	578	53.3%	96.6%	2.4%	0	0	-	0.0%	0.0%
Dementia	2,985	77,575	729	11	96.3%	24.4%	0.01%	2,388	583	71.4%	80.0%	7.5%
Hypo-thyroidism	2,030	25,760	2,029	258	9.6%	99.9%	1.0%	2,029	1,065	10.7%	99.9%	4.1%

TP=true positive; FP=false positive; FDR=false discovery rate; T2DM=type 2 diabetes mellitus.

medRxiv preprint doi: https://doi.org/10.1101/2023.12.19.23300230; this version posted December 19, 2023. The copyright holder for this preprint **(which was not certified by peer review)** is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.

It is made available under a CC-BY-NC 4.0 International license .

DISCUSSION

EHR phenotyping is a critical area of modern observational clinical research, yet it commonly demands substantial resources. In this study, we explored the effectiveness of LLMs in creating preliminary versions of computable phenotyping algorithms, with the ultimate goal of streamlining the EHR phenotyping process.

Not all LLMs we tested were well-suited for phenotyping. ChatGPT-4 and ChatGPT-3.5 significantly outperformed both Claude 2 and Bard in their ability to generate executable and accurately SQL-formatted phenotyping algorithms. One of the reasons for this discrepancy was Claude 2's tendency to represent concepts using numerical concept codes, without specifying what clinical criteria these concept codes were intended to capture. Five of the algorithms generated by Bard did not follow the OMOP CDM. Four of these algorithms referenced columns that do not exist in the OMOP CDM. These corresponded to “patient_id” for both prompting strategies of dementia, “observation_fact” and “measurement_fact” for the α-prompting strategies of T2DM, and “occurrence_age” for the β-prompting strategy of T2DM. Additionally, two of the algorithms searched for concepts in the wrong tables. The α-prompting algorithm for dementia attempted to use the “person” table to extract all concepts, including diagnosis codes and medications, which would not be found in this table, whereas the β-prompting algorithm for hypothyroidism looked for signs and symptoms in the Measurement table. Given the poor performance of the phenotyping algorithms generated by Claude 2 and Bard, we focused the remaining analysis on phenotyping algorithms generated by ChatGPT-4 and ChatGPT-3.5.

Both ChatGPT models were able to follow instructions and identify relevant concepts for the three selected phenotypes. The phenotyping algorithms generated by these models contained reasonably accurate diagnosis codes, related lab tests, and key medications. The concepts identified largely overlapped with the ones used by domain experts. We found that ChatGPT-4 generally demonstrated slightly superior performance compared to ChatGPT-3.5 by identifying more medications and providing more appropriate thresholds for lab values (**Table 1**). Moreover, the ChatGPT-generated algorithms were able to identify additional potentially useful criteria, including a variety of symptoms and clinical signs for each phenotype, which have not usually been incorporated in any eMERGE algorithms.

medRxiv preprint doi: https://doi.org/10.1101/2023.12.19.23300230; this version posted December 19, 2023. The copyright holder for this preprint **(which was not certified by peer review)** is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.

It is made available under a CC-BY-NC 4.0 International license .

ChatGPT generated phenotyping algorithms with incorrect criteria that needed manual corrections to ensure their functionality. For example, both ChatGPT-4 and ChatGPT-3.5 erroneously considered the ICD-10 code F00 as a relevant diagnosis code for dementia, whereas our prompts specifically required ICD-10-CM codes. Additionally, ChatGPT occasionally selected inappropriately broad ICD-9-CM codes, such as 250* for T2DM, inadvertently encompassing diagnoses related to other types of diabetes. Also, ChatGPT missed some key ICD codes and medications. For example, while the query correctly identified patients with ICD-10-CM code G30 as dementia cases, ChatGPT overlooked patients with more specific codes such as G30.0, G30.1, G30.8, and G30.9. The ChatGPT algorithms generally generated a shorter list of medications compared to their corresponding eMERGE algorithms, with the exception of the ChatGPT-4 algorithm for hypothyroidism. In this case, ChatGPT-4 interpreted the hypothyroidism phenotype more broadly to include not only endogenous causes of hypothyroidism (as in the eMERGE algorithm) but also exogenous causes (not included in the eMERGE algorithm). As a result, the medications specified by the ChatGPT-4 hypothyroidism algorithm included both drugs used for treatment of hypothyroidism and drugs with potential to cause hypothyroidism (i.e., lithium, amiodarone, Lithobid, Cordarone, Nexterone, Pacerone). Moreover, certain thresholds set for lab values were inaccurate, such as “fasting blood glucose ≥6.5” for T2DM.

As noted above, compared to the eMERGE algorithms, ChatGPT introduced some new concepts, including signs and symptoms such as fatigue, cold intolerance, and weight gain. Many of these concepts are infrequently used by domain experts in algorithm generation due to their low specificity. Including these low-specificity concepts might not substantially enhance the recall of the algorithms and could potentially diminish the algorithm’s PPV. Despite this, ChatGPT showed promise in identifying noteworthy concepts, such as the Mini-Mental State Examination and Montreal Cognitive Assessment for dementia.

Finally, the LLMs exhibited immature capability in organizing phenotyping criteria with the proper logic. The SQL queries generated by ChatGPT were predominantly characterized by a single logical operator (AND or OR), resulting in phenotyping algorithms that were either excessively restrictive or overly broad.

Limitations and future work

medRxiv preprint doi: https://doi.org/10.1101/2023.12.19.23300230; this version posted December 19, 2023. The copyright holder for this preprint **(which was not certified by peer review)** is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.

It is made available under a CC-BY-NC 4.0 International license .

Several limitations need to be highlighted as potential opportunities for future improvement. First, this pilot study was limited in scope and did not investigate all the possible tools, options, and capabilities of the LLMs listed here. We are reporting initial observations from using the most basic and widely accessible version of these LLMs, i.e., the chat box interface with default settings available through the websites of each of the LLMs. Using an API would allow for selection of additional parameters, which may affect performance. Second, as a pilot study, this research focused on prompting LLMs to generate algorithms for identifying phenotype cases. The capability of LLMs in generating algorithms to identify controls also needs to be evaluated. Third, we tested solely on proprietary models and their default configurations. It is important to assess both proprietary models and the leading open-source models (e.g., Llama 2), especially when they are enhanced with fine-tuning capabilities and knowledge integration. Fourth, our design of prompts did not consider the further optimization of the execution efficiency of the SQL queries. Consequently, LLMs often produced SQL queries with suboptimal query structures, leading to slow execution in our database and additional human efforts to refine the query. Fifth, this study only considered phenotypes for three common diseases. Phenotyping rare diseases may present different challenges particularly when there is limited relevant online content for a particular disease. Sixth, the consistency of responses generated by ChatGPT varies over time, warranting consideration for potential future investigations. New LLMs are being rapidly deployed, and our future efforts will involve exploring alternative advanced models, such as ResearchGPT and Google Gemini. Additionally, we will be delving into more refined control and customization in the generation process through prompt engineering to achieve desired performance levels.

medRxiv preprint doi: https://doi.org/10.1101/2023.12.19.23300230; this version posted December 19, 2023. The copyright holder for this preprint **(which was not certified by peer review)** is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.

It is made available under a CC-BY-NC 4.0 International license .

CONCLUSION

ChatGPT models are capable of producing phenotyping algorithm drafts that align with a standard CDM and can be executed with reasonable modifications. These models excel in identifying relevant clinical concepts that can serve as valuable initial components for defining phenotypes. Nevertheless, LLM-generated phenotyping algorithms still necessitate manual modifications using clinical and informatics knowledge.

medRxiv preprint doi: https://doi.org/10.1101/2023.12.19.23300230; this version posted December 19, 2023. The copyright holder for this preprint **(which was not certified by peer review)** is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.

It is made available under a CC-BY-NC 4.0 International license .

AUTHOR CONTRIBUTIONS

W.Q.W. conceived and supervised this study. C.Y. designed the interactions with the large language models (LLM), collected data, and analyzed the results. W.Q.W. and H.H.O. implemented the algorithms produced by LLMs. M.E.G., V.E.K., and W.Q.W. performed expert review for the produced algorithms. M.E.G., C.Y., H.H.O., V.E.K., and M.S.K. drafted the paper. W.Q.W., B.A.M., W.C.S., A.L.D., Q.P.F., and C.M.S. critically revised the paper. D.M.R. and J.F.P. reviewed the paper and provided important intellectual content. All authors approved this study.

medRxiv preprint doi: https://doi.org/10.1101/2023.12.19.23300230; this version posted December 19, 2023. The copyright holder for this preprint **(which was not certified by peer review)** is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.

It is made available under a CC-BY-NC 4.0 International license .

FUNDING

This work was supported by R01GM139891, R01AG069900, F30AG080885, T32GM007347, K01HL157755, and U01HG011181.

medRxiv preprint doi: https://doi.org/10.1101/2023.12.19.23300230; this version posted December 19, 2023. The copyright holder for this preprint **(which was not certified by peer review)** is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.

It is made available under a CC-BY-NC 4.0 International license .

CONFLICTS OF INTEREST

All authors have no competing interests to declare.

medRxiv preprint doi: https://doi.org/10.1101/2023.12.19.23300230; this version posted December 19, 2023. The copyright holder for this preprint **(which was not certified by peer review)** is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.

It is made available under a CC-BY-NC 4.0 International license .

CODE AVAILABILITY

Source code for results and LLM-generated SQL-formatted algorithms are shared at https://github.com/The-Wei-Lab/LLM-Phenotyping-2024.

medRxiv preprint doi: https://doi.org/10.1101/2023.12.19.23300230; this version posted December 19, 2023. The copyright holder for this preprint **(which was not certified by peer review)** is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.

It is made available under a CC-BY-NC 4.0 International license .

REFERENCE

1. Wei WQ, Denny JC. Extracting research-quality phenotypes from electronic health records to support precision medicine. *Genome Med*. 2015;7(1):41.
2. Banda JM, Seneviratne M, Hernandez-Boussard T, Shah NH. Advances in electronic phenotyping: From rule-based definitions to machine learning models. *Annu Rev Biomed Data Sci*. 2018;1(1):53-68.
3. Pacheco JA, Rasmussen LV, Wiley K Jr, et al. Evaluation of the portability of computable phenotypes with natural language processing in the eMERGE network. *Sci Rep*. 2023;13(1):1971.
4. Newton KM, Peissig PL, Kho AN, et al. Validation of electronic medical record-based phenotyping algorithms: results and lessons learned from the eMERGE network. *J Am Med Inform Assoc*. 2013;20(e1):e147-54.
5. Zheng NS, Feng Q, Kerchberger VE, et al. PheMap: a multi-resource knowledge base for high-throughput phenotyping within electronic health records. *J Am Med Inform Assoc*. 2020;27(11):1675-1687.
6. Wu P, Gifford A, Meng X, et al. Mapping ICD-10 and ICD-10-CM codes to phecodes: Workflow development and initial evaluation. *JMIR Med Inform*. 2019;7(4):e14325.
7. Grabowska ME, Van Driest SL, Robinson JR, et al. Developing and evaluating pediatric phecodes (Peds-phecodes) for high-throughput phenotyping using electronic health records. *medRxiv*. Published online 2023. doi:10.1101/2023.08.22.23294435.
8. Zhang Y, Cai T, Yu S, et al. High-throughput phenotyping with electronic medical record data using a common semi-supervised approach (PheCAP). *Nat Protoc*. 2019;14(12):3426-3444.
9. Clusmann J, Kolbinger FR, Muti HS, et al. The future landscape of large language models in medicine. *Commun Med (Lond)*. 2023;3(1):1-8.
10. OpenAI. GPT-4 Technical Report. Preprint at https://doi.org/10.48550/arXiv.2303.08774 (2023).
11. OpenAI. Introducing ChatGPT. https://openai.com/blog/chatgpt (2022).
12. Introducing Claude. *Anthropic* https://www.anthropic.com/index/introducing-claude (2023).
13. Pichai, S. Google AI updates: Bard and new AI features in Search. https://blog.google/technology/ai/bard-google-ai-search-updates/ (2023).
14. Gottesman O, Kuivaniemi H, Tromp G, Faucett WA, Li R, Manolio TA, Sanderson SC, Kannry J, Zinberg R, Basford MA, Brilliant M. The electronic medical records and genomics (eMERGE) network: past, present, and future. *Genetics in Medicine.* 2013 Oct;15(10):761-71.
15. McCarty CA, Chisholm RL, Chute CG, Kullo IJ, Jarvik GP, Larson EB, Li R, Masys DR, Ritchie MD, Roden DM, Struewing JP. The eMERGE Network: a consortium of biorepositories linked to electronic medical records data for conducting genomic studies. *BMC medical genomics.* 2011 Dec;4:1-1.
16. Omiye JA, Lester JC, Spichak S, Rotemberg V, Daneshjou R. Large language models propagate race-based medicine. *NPJ Digit Med*. 2023;6(1):195.
17. Kho AN, Hayes MG, Rasmussen-Torvik L, Pacheco JA, Thompson WK, Armstrong LL, et al. Use of diverse electronic medical record systems to identify genetic risk for type 2 diabetes within a genome-wide association study. *J Am Med Inform Assoc*. 2012;19(2):212–8.
18. Jennifer Pacheco and Will Thompson. Northwestern University. Type 2 diabetes mellitus. *Phekb.org*. 2012. Available from: https://phekb.org/phenotype/type-2-diabetes-mellitus.
19. Chris Carlson. Group Health Cooperative. Dementia. *PheKB.org.* 2012. Available from: https://phekb.org/phenotype/10.
20. Josh Denny. Group Health Cooperative, Marshfield Clinic Research Foundation, Mayo Medical School College of Medicine, Northwestern University, Vanderbilt University. Hypothyroidism. *PheKB.org*. 2012. Available from: https://phekb.org/phenotype/14.

medRxiv preprint doi: https://doi.org/10.1101/2023.12.19.23300230; this version posted December 19, 2023. The copyright holder for this preprint **(which was not certified by peer review)** is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.

It is made available under a CC-BY-NC 4.0 International license .

21. Denny JC, Crawford DC, Ritchie MD, Bielinski SJ, Basford MA, Bradford Y, et al. Variants near FOXE1 are associated with hypothyroidism and other thyroid conditions: using electronic medical records for genome- and phenome-wide studies. *Am J Hum Genet.* 2011;89(4):529–42.
22. OMOP Common Data Model. Github.io. https://ohdsi.github.io/CommonDataModel/.
23. OHDSI – observational health data sciences and informatics. Ohdsi.org. https://www.ohdsi.org/.
24. Wei J, Wang X, Schuurmans D, Bosma M, Xia F, Chi E, Le QV, Zhou D. Chain-of-thought prompting elicits reasoning in large language models. *Proc of the 37th Conference on Neural Information Processing Systems.* 2022 Dec 6;35:24824-37.
25. Madaan A, Tandon N, Gupta P, Hallinan S, Gao L, Wiegreffe S, et al. Self-refine: Iterative refinement with Self-feedback [Internet]. arXiv [cs.CL]. 2023. Available from: http://arxiv.org/abs/2303.17651.
26. Pan L, Saxon M, Xu W, Nathani D, Wang X, Wang WY. Automatically Correcting Large Language Models: Surveying the landscape of diverse self-correction strategies [Internet]. arXiv [cs.CL]. 2023. Available from: http://arxiv.org/abs/2308.03188.
27. Samwald M, Praas R, Hebenstreit K. Towards unified objectives for self-reflective AI. SSRN Electron J [Internet]. 2023; Available from: http://dx.doi.org/10.2139/ssrn.4446991.
28. Vanbelle S. A new interpretation of the weighted kappa coefficients. *Psychometrika*. 2016 Jun;81(2):399-410.
29. Woolson RF. Wilcoxon signed-rank test. *Wiley encyclopedia of clinical trials.* 2007 Mar 9:1-3.
30. Wan NC, Yaqoob AA, Ong HH, Zhao J, Wei W-Q. Evaluating resources composing the PheMAP knowledge base to enhance high-throughput phenotyping. *J Am Med Inform Assoc*. 2023;30(3):456–65.
31. Zheng NS, Feng Q, Kerchberger VE, Zhao J, Edwards TL, Cox NJ, et al. PheMap: a multi-resource knowledge base for high-throughput phenotyping within electronic health records. *J Am Med Inform Assoc*. 2020;27(11):1675–87.
32. Landis JR, Koch GG. The measurement of observer agreement for categorical data. *Biometrics*. 1977;33(1):159–74.

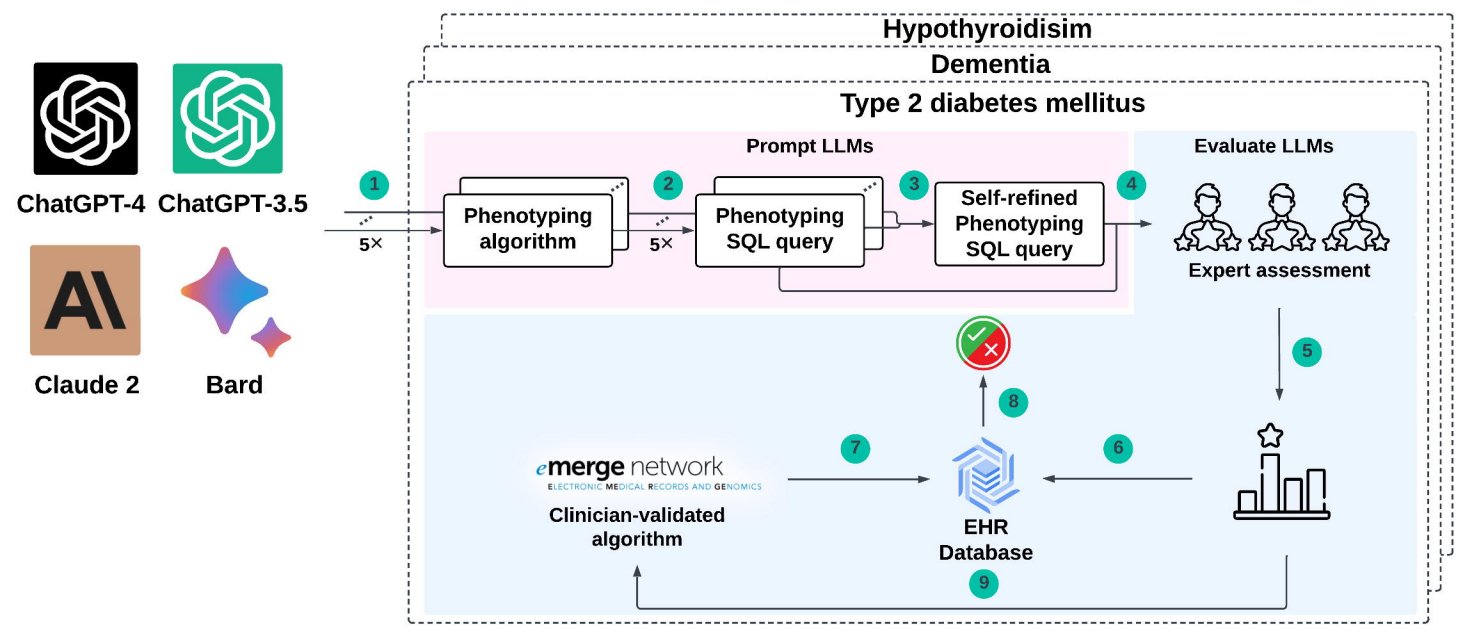
medRxiv preprint doi: https://doi.org/10.1101/2023.12.19.23300230; this version posted December 19, 2023. The copyright holder for this preprint **(which was not certified by peer review)** is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.

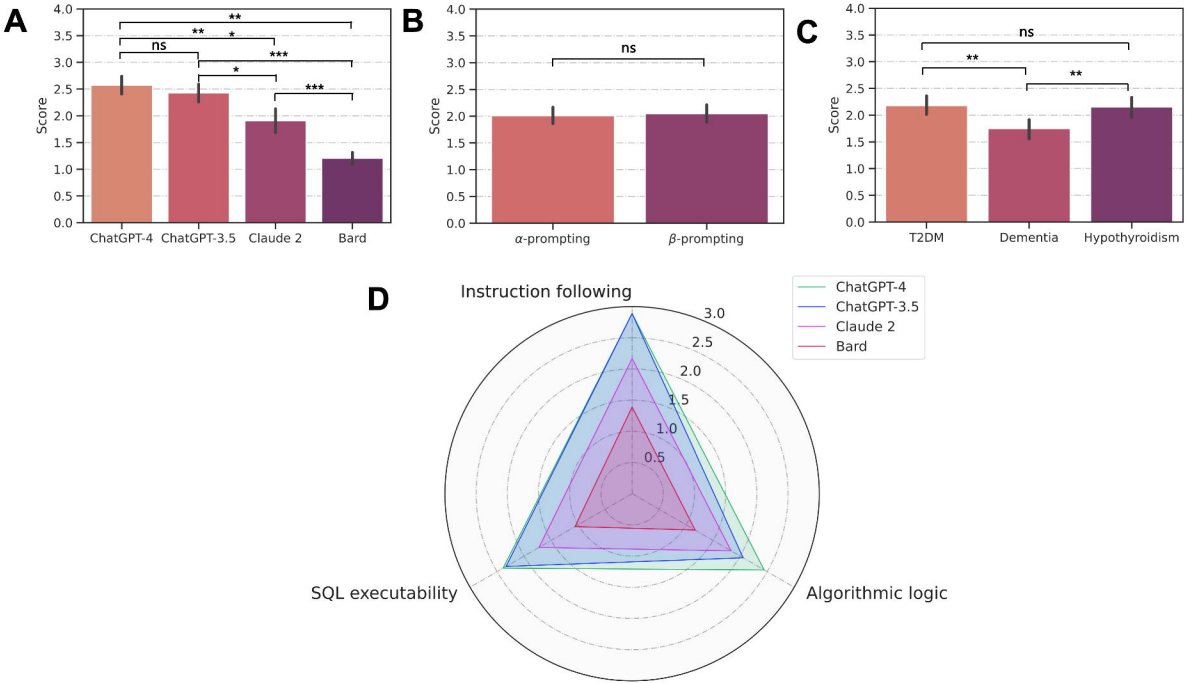
It is made available under a CC-BY-NC 4.0 International license .

FIGURE LEGENDS

Figure 1. An architectural overview of the study pipeline.

Figure 2. A comparative analysis based on expert evaluations focusing on A) four large language models, B) two prompting strategies, C) three phenotypes, and D) three individual evaluation axes. Numeric scores of 3, 2, and 1 correspond to expert assessments of “Good”, “Medium”, and “Poor”, respectively. ***, **, and * denote $p<0.001$, $p<0.01$, and $p<0.05$, respectively. ns=not significant.


ChatGPT-4
ChatGPT-3.5
Claude 2
Bard
Hypothyroidisim
Dementia
Type 2 diabetes mellitus
Prompt LLMs
Evaluate LLMs
1
2
3
4
5
6
7
8
9
5×
5×
Phenotyping algorithm
Phenotyping SQL query
Self-refined Phenotyping SQL query
Expert assessment
emerge network
ELECTRONIC MEDICAL RECORDS AND GENOMICS
Clinician-validated algorithm
EHR Database


A
Score
ChatGPT-4
ChatGPT-3.5
Claude 2
Bard
ns
**
*
***
B
ns
α-prompting
β-prompting
C
ns
**
T2DM
Dementia
Hypothyroidism
D
Instruction following
SQL executability
Algorithmic logic
ChatGPT-4
ChatGPT-3.5
Claude 2
Bard