NEJM AI

NEJM AI 2024; 1 (5)
DOI: 10.1056/AIdbp2300040

DATASETS, BENCHMARKS, AND PROTOCOLS

Large Language Models Are Poor Medical Coders — Benchmarking of Medical Code Querying

Ali Soroush, M.D., M.S.,[1,2,3] Benjamin S. Glicksberg, Ph.D.,[1,2] Eyal Zimlichman, M.D., M.Sc.,[4,5] Yiftach Barash, M.D., M.Sc.,[5,6] Robert Freeman, R.N., M.S.N., N.E.-B.C.,[7] Alexander W. Charney, M.D., Ph.D.,[1,2] Girish N Nadkarni, M.D., M.P.H.,[1,2] and Eyal Klang, M.D.[1,2,5,8]

Received: July 16, 2023; Revised: February 4, 2024; Accepted: March 1, 2024; Published: April 19, 2024

Abstract

BACKGROUND Large language models (LLMs) have attracted significant interest for automated clinical coding. However, early data show that LLMs are highly error-prone when mapping medical codes. We sought to quantify and benchmark LLM medical code querying errors across several available LLMs.

METHODS We evaluated GPT-3.5, GPT-4, Gemini Pro, and Llama2-70b Chat performance and error patterns when querying medical billing codes. We extracted 12 months of unique International Classification of Diseases, 9th edition, Clinical Modification (ICD-9-CM), International Classification of Diseases, 10th edition, Clinical Modification (ICD-10-CM), and Current Procedural Terminology (CPT) codes from the Mount Sinai Health System electronic health record (EHR). Each LLM was provided with a code description and prompted to generate a billing code. Exact match accuracy and other performance metrics were calculated. Nonexact matches were analyzed using descriptive metrics and standardized measures of text and code similarity, including METEOR score, BERTScore, and cui2-vec cosine similarity. We created and applied a CodeSTS manual similarity grading system to 200 randomly selected codes weighted by EHR code frequency. Using CodeSTS scores, we identified correct "equivalent" or "generalized" generated codes.

RESULTS A total of 7697 ICD-9-CM, 15,950 ICD-10-CM, and 3673 CPT codes were extracted. GPT-4 had the highest exact match rate (ICD-9-CM: 45.9%; ICD-10-CM: 33.9%; CPT: 49.8%). Among incorrectly matched codes, GPT-4 generated the most equivalent codes (ICD-9-CM: 7.0%; ICD-10-CM: 10.9%), and GPT-3.5 generated the most generalized but correct codes (ICD-9-CM: 29.9%; ICD-10-CM: 18.5%). Extracted code frequency, shorter codes, and shorter code descriptions were associated ($P<0.05$) with higher exact match rates in nearly all analyses.

CONCLUSIONS All tested LLMs performed poorly on medical code querying, often generating codes conveying imprecise or fabricated information. LLMs are not appropriate

Drs. Nadkarni and Klang contributed equally to this article.

The author affiliations are listed at the end of the article.

Dr. Soroush can be contacted at ali.soroush@mountsinai.org or at 1 Gustave L. Levy Place, Box 1003, New York, NY 10029.

for use on medical coding tasks without additional research. (Funded by the AGA Research Foundation and National Institutes of Health.)

Introduction

The International Classification of Diseases (ICD) terminology is the most widely used administrative coding system and provides a standardized representation of medical diagnoses.[1-3] In the United States, the "CM" (clinical modification) of this terminology plays a critical role in clinical recordkeeping, public health surveillance, research, and billing.[2] Current Procedural Terminology (CPT) codes are analogously used for procedural billing.[4] Automating the extraction of medical codes from unstructured clinical text has been a longstanding goal of medical natural language processing (NLP) research.[5] Thus far, automated clinical coding systems require significant engineering resources to deploy and demonstrate insufficient accuracy, leaving most health care systems to rely on manual coders.[5]

Large language models (LLMs) are deep learning models trained on extensive textual data, capable of generating text output.[6] LLMs have shown remarkable text processing and reasoning capabilities, suggesting that they could automate key administrative tasks.[7-10] However, even the best LLMs extract fewer correct ICD-10-CM codes and generate more incorrect codes from clinical text than smaller fine-tuned language models.[11] We sought to benchmark baseline LLM performance on medical code querying across several available LLMs, including GPT-3.5, GPT-4, Gemini Pro, and LLaMa-70b Chat. We aimed to identify potential mechanisms to improve code mapping performance. Using a systematic and automated approach whereby we instruct the LLM to generate a code when provided the code's description, we aimed to characterize the medical coding capabilities of current LLMs in sufficient detail to guide additional research.

Methods

CODE EXTRACTION FROM MOUNT SINAI DATA WAREHOUSE

We extracted unique primary International Classification of Diseases, 9th edition, Clinical Modification (ICD-9-CM), International Classification of Diseases, 10th edition, Clinical Modification (ICD-10-CM), and CPT billing codes from Mount Sinai Health System electronic health records (EHRs) collected during the time periods corresponding with the most recent Centers for Medicare & Medicaid Services (CMS) billing code lists: October 1, 2014, to September 30, 2015, for ICD-9-CM and October 1, 2022, to September 30, 2023, for ICD-10-CM and CPT. We used the UMLS (Unified Medical Language System) REST API (representational state transfer application programming interface; UMLS Metathesaurus version: 2023AB release) to obtain the preferred description for each code.[12]

LLM CODE GENERATION

We utilized GPT-3.5 Turbo (March 2023, June 2023, and November 2023 versions), GPT-4 (March 2023, June 2023, and November 2023 versions), Gemini Pro, and Llama2-70b Chat (Table S1 in the Supplementary Appendix) to assess medical code querying capabilities.[13-16] The APIs for these models were accessed between December 26 and 27, 2023. Our primary task involved prompting the models to generate a code when given the preferred code description of an extracted code. We constructed a standardized prompt for this task (see Supplementary Methods).[17] To standardize our LLM API calls, we used LangChain (version 0.350). All models were set to a temperature of 0.2 and 50 maximum output tokens. We selected the lowest temperature that produced valid output across all models to reduce response variability. We tested temperatures of 0.2, 0.4, 0.6, 0.8, and 1.0 on a subset of 100 codes and did not notice meaningful differences in overall accuracy. We confirmed that no LLM responses were truncated due to the maximum output tokens. We used the UMLS REST API to obtain the preferred code description for each generated code.[12]

MANUAL ASSESSMENT OF CONCEPT SIMILARITY

To manually assess concept similarity, we developed CodeSTS, an adaptation of MedSTS (medical semantic textual similarity)[18] that quantifies differences between original and generated code descriptions. We applied weighted sampling based on the frequency of EHR codes to randomly select 220 codes from each list of unique ICD-9-CM, ICD-10-CM, and CPT codes. Twenty of these codes were jointly reviewed by two physicians (E.K. and A.S.) to develop a consistent ruleset for assigning CodeSTS scores (Table S2). We (E.K. and A.S.) independently scored the dissimilar descriptions for the remaining 200 codes and used the average scores for our analyses. We used Cohen's

kappa to assess interrater reliability. There was a moderate degree of interrater reliability between the two reviewers (ICD-9-CM: 0.541; ICD-10-CM: 0.596; CPT: 0.458). We also assessed concordance between CodeSTS and the calculated similarity scores (METEOR score,[19] BERTScore,[20] and cui2vec[21] cosine similarity) but found poor concordance (<0.40) across all metrics.

PERFORMANCE EVALUATION

The rate of exact code matches was calculated as an overall metric of model performance. Within the manual subset, we also determined the number of equivalent codes ($4 \geq$ CodeSTS > 5). To assess the similarity in text and meaning between the original and generated codes holistically, we utilized METEOR,[19] BERTScore,[20] and cui2vec[21] cosine similarity. Each automated metric assesses similarity in a complementary fashion (Table S3). In the manual subset, we also utilized CodeSTS scores to evaluate similarity.

ERROR ANALYSIS

To identify code generation error patterns, we analyzed the generated codes that were not exact matches. We assessed the rate of valid codes and fabricated (nonvalid) codes. Valid codes were defined as codes present from the UMLS Metathesaurus.[12] We also assessed the rate of billable and nonbillable ICD codes. Nonbillable codes represent concepts that are too vague for billing purposes. We defined these codes as the subset of valid codes not present in the most recent CMS list of acceptable billing codes (ICD-9-CM: 2014; ICD-10-CM: 2023).[22] All CPT codes are "billable," so no distinction was made for CPT codes. Because ICD codes become more granular with each successive digit, we measured the longest sequence of matched digits as an additional means of assessing partial accuracy. To assess the structural patterns of each model's code generation, we assessed the rate of correctly matched code length and correct digit-level matches. We qualitatively and quantitatively assessed the relationships between exact match rates and code length, code description, and log-transformed EHR code frequency using histograms and correlation coefficients, respectively. We used tornado plots to visualize the relationships between exact match and length, description length, and log-transformed EHR code frequency.

In the manual subset, we used the CodeSTS scores to identify codes that were correct, either because they were equivalent (CodeSTS ≥ 4) or generalized ($3 <$ CodeSTS ≤ 4). Calculated metrics (METEOR score,[19] BERTScore,[20] and cui2vec[21] cosine similarity) and the manual CodeSTS score were also assessed. CodeSTS score distributions for the incorrect codes were described using histograms.

STATISTICAL ANALYSIS

We reported the median and interquartile range for code frequency in the Mount Sinai Health System for 1 year. We summarized the ICD code querying performance for each model using descriptive statistics. We calculated mean values for exact match, equivalent match, generalized match, valid codes, billable codes, nonbillable codes, fabricated codes, matched length, matched digits, longest sequence of correct digits matched, CodeSTS score, METEOR score, BERTScore, and cui2vec cosine similarity. We reported median values for code generation frequency. We calculated 95% confidence intervals for all values using bootstrapping with 10,000 samples. We calculated point-biserial correlation coefficients and their P values for the relationships between each model's performance and code frequency, code length, and description length using an alpha of 0.05. We coded all analyses in Python (Version 3.11.5).

Results

CODE DATASETS

We extracted 7697 unique ICD-9-CM codes from the EHR, with a median code frequency of 18.0 (interquartile range: 4.0 to 89.0, maximum: 102,072). We extracted 15,950 unique ICD-10-CM codes, with a median frequency of 9.0 (interquartile range: 2.0 to 45.0, maximum: 140,560). We extracted 3673 unique CPT codes with a median code frequency of 8.0 (interquartile range: 2.0 to 97.0, maximum: 1,304,462). A total of 200 unique codes, weighted by frequency of use within the EHR, were randomly selected for each code system. The manually reviewed ICD-9-CM, ICD-10-CM, and CPT data had median code frequencies of 1511.5 (interquartile range: 478.0 to 4603.0), 1735.0 (interquartile range: 521.5 to 6191.25), and 15,977.5 (6080.0 to 64,226.0), respectively. The code frequency distributions for the full code sets and manual subsets are shown in Figure 1.

CODE GENERATION PERFORMANCE EVALUATION

Code generation performance for the full dataset varied across the different coding systems and models (Table 1). GPT-4 (November [Nov]) had the highest exact match rates

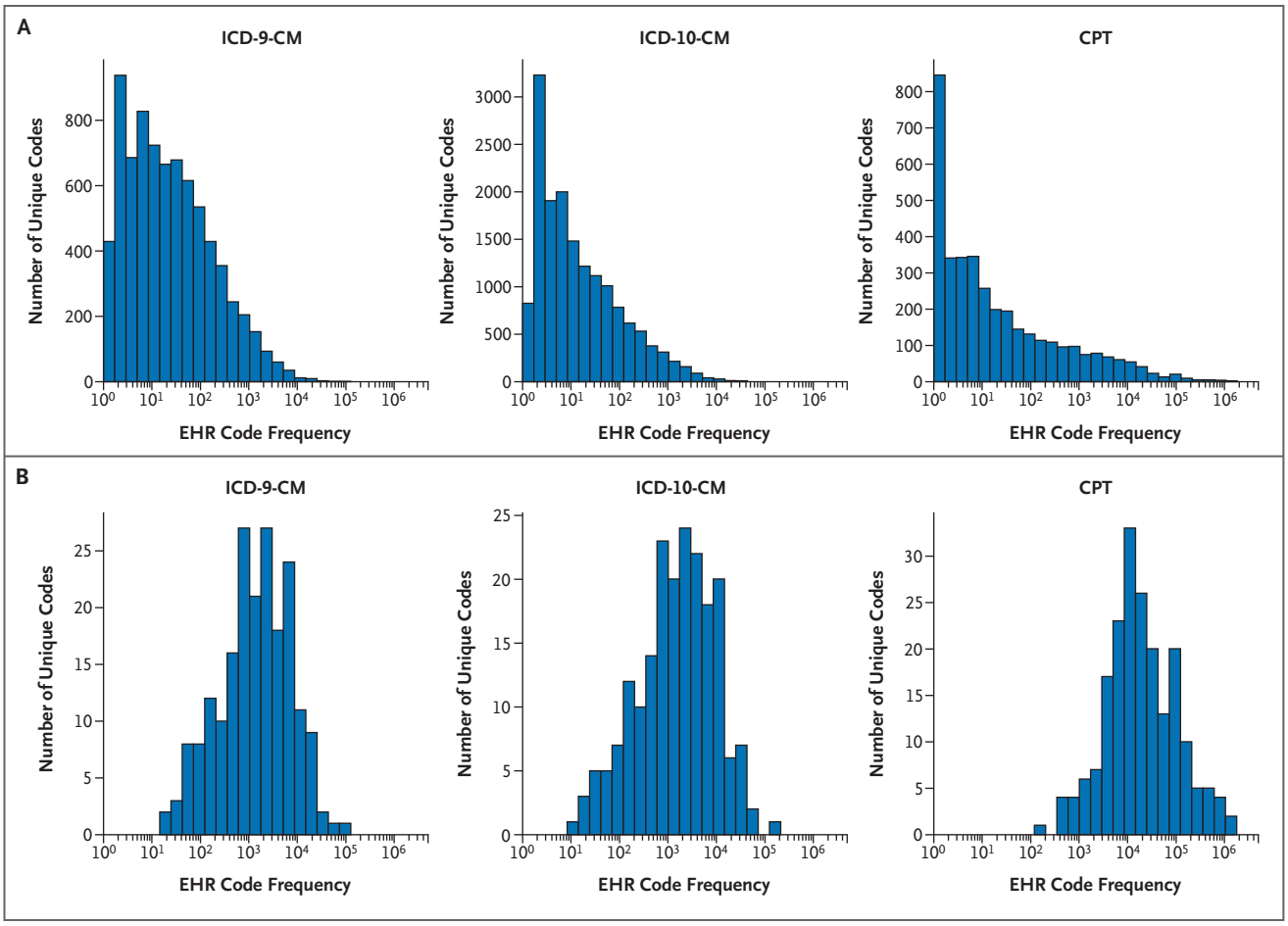

Figure 1. Frequency of Codes in the Electronic Health Record (EHR).

Panel A shows all extracted codes. Panel B shows a manually reviewed subset of 200 codes per system, randomly selected with weighted sampling based on the frequency of electronic health record codes. CPT indicates Current Procedural Terminology; ICD-9-CM, International Classification of Diseases, 9th edition, Clinical Modification; and ICD-10-CM, International Classification of Diseases, 10th edition, Clinical Modification.

(ICD-9-CM: 45.9%; ICD-10-CM: 33.9%; CPT: 49.8%), and Llama2-70b Chat scored the lowest (ICD-9-CM: 1.2%; ICD-10-CM: 1.5%; CPT: 2.6%). When excluding GPT-4, GPT-3.5 (Nov) had the next best match rates (ICD-9-CM: 28.9%; ICD-10-CM: 18.2%; CPT: 31.9%), followed by Gemini Pro (ICD-9-CM: 10.7%; ICD-10-CM: 4.8%; CPT: 11.4%) and Llama2-70b Chat (ICD-9-CM: 1.2%; ICD-10-CM: 1.5%; CPT: 2.6%). Both GPT-4 and GPT-3.5 Turbo demonstrated improved exact match performance with each successive model. At the code system level, ICD-9-CM and CPT codes generally had more exact matches than ICD-10-CM. The only exception was Llama2-70b Chat, which had the lowest match rate with ICD-9-CM.

We assessed textual similarity between each pair of original and generated code descriptions using METEOR scores and BERTScores. GPT-4 (Nov) consistently scored highest and Llama2-70b Chat consistently scored worst across all code systems and scores. Gemini Pro and the best-scoring GPT-3.5 Turbo model had scores between these two extremes. To assess the conceptual similarity between the original and generated codes, we calculated cui2vec cosine similarity scores and observed a similar pattern. GPT-4 (Nov) scored highest (ICD-9-CM: 0.843; ICD-10-CM: 0.733), followed by GPT-3.5 Turbo (best ICD-9-CM score: 0.765; best ICD-10-CM score: 0.566), Gemini Pro (ICD-9-CM: 0.641; ICD-10-CM: 0.414), and Llama2-70b Chat (ICD-9-CM: 0.418; ICD-10-CM: 0.287).

When evaluating the manual data (Table 2), patterns of model performance were unchanged. However, there was an overall increase in exact match rates and conceptual

Table 1. ICD-9-CM, ICD-10-CM, and CPT Code Generation Overall Performance Metrics, Full Code Set.*

Coding System	Metric	GPT-3.5 Turbo (March)†	GPT-3.5 Turbo (June)†	GPT-3.5 Turbo (Nov)†	GPT-4 (March)†	GPT-4 (June)†	GPT-4 (Nov)†	Gemini Pro†	Llama2-70b Chat†
ICD-9-CM (n=7697)	Exact match, % (95% CI)	26.6% (25.6%–27.6%)	26.7% (25.7%–27.7%)	28.9% (27.9%–29.9%)	42.3% (41.2%–43.4%)	44.1% (43.0%–45.2%)	45.9% (44.8%–47.0%)	10.7% (10.0%–11.4%)	1.2% (1.0%–1.5%)
	cui2vec cosine similarity, mean (95% CI)	0.747 (0.741–0.753)	0.750 (0.744–0.756)	0.765 (0.760–0.771)	0.833 (0.828–0.838)	0.837 (0.832–0.842)	0.843 (0.838–0.848)	0.641 (0.635–0.648)	0.418 (0.409–0.428)
	METEOR score, mean (95% CI)	0.415 (0.406–0.424)	0.414 (0.405–0.422)	0.437 (0.428–0.445)	0.564 (0.555–0.573)	0.579 (0.569–0.588)	0.593 (0.585–0.602)	0.255 (0.247–0.262)	0.100 (0.094–0.106)
	BERTScore, mean (95% CI)	0.857 (0.855–0.860)	0.856 (0.854–0.859)	0.863 (0.861–0.866)	0.899 (0.896–0.901)	0.903 (0.901–0.906)	0.907 (0.904–0.909)	0.812 (0.809–0.814)	0.749 (0.747–0.751)
ICD-10-CM (n=15,950)	Exact match, % (95% CI)	17.1% (16.5%–17.7%)	17.8% (17.2%–18.4%)	18.2% (17.6%–18.8%)	27.5% (26.8%–28.1%)	28.4% (27.7%–29.1%)	33.9% (33.2%–34.6%)	4.8% (4.5%–5.1%)	1.5% (1.4%–1.7%)
	cui2vec cosine similarity, mean (95% CI)	0.571 (0.564–0.577)	0.576 (0.570–0.583)	0.566 (0.559–0.572)	0.669 (0.663–0.675)	0.680 (0.673–0.685)	0.733 (0.728–0.739)	0.414 (0.406–0.421)	0.287 (0.280–0.294)
	METEOR score, mean (95% CI)	0.399 (0.393–0.405)	0.405 (0.399–0.410)	0.400 (0.394–0.406)	0.510 (0.504–0.516)	0.522 (0.516–0.528)	0.581 (0.575–0.587)	0.250 (0.245–0.254)	0.129 (0.125–0.132)
	BERTScore, mean (95% CI)	0.866 (0.864–0.868)	0.870 (0.868–0.871)	0.866 (0.864–0.868)	0.899 (0.897–0.900)	0.902 (0.901–0.904)	0.918 (0.917–0.920)	0.824 (0.822–0.826)	0.774 (0.773–0.776)
CPT (n=3673)	Exact match, % (95% CI)	28.4% (27.0%–29.9%)	26.2% (24.7%–27.6%)	31.9% (30.4%–33.4%)	44.0% (42.4%–45.6%)	42.6% (41.0%–44.2%)	49.8% (48.2%–51.5%)	11.4% (10.3%–12.4%)	2.6% (2.1%–3.1%)
	METEOR score, mean (95% CI)	0.461 (0.448–0.474)	0.433 (0.421–0.446)	0.495 (0.482–0.507)	0.596 (0.583–0.609)	0.586 (0.573–0.599)	0.655 (0.642–0.667)	0.295 (0.284–0.306)	0.182 (0.172–0.192)
	BERTScore, mean (95% CI)	0.868 (0.864–0.871)	0.859 (0.855–0.863)	0.878 (0.874–0.882)	0.904 (0.901–0.908)	0.901 (0.897–0.904)	0.921 (0.918–0.925)	0.816 (0.813–0.820)	0.770 (0.766–0.773)

* CI indicates confidence interval; CPT, Current Procedural Terminology; ICD-9-CM, International Classification of Diseases, 9th edition, Clinical Modification; ICD-10-CM, International Classification of Diseases, 10th edition, Clinical Modification; and Nov, November.

† The application programming interface was accessed between December 26 and 27, 2023.

scores across all models and code systems. The equivalent match rate (CodeSTS score $\geq$ 4) was highest for GPT-4 (Nov) (ICD-9-CM: 80.0%; ICD-10-CM: 78.0%; CPT: 83.5%) and lowest for Llama2-70b Chat (ICD-9-CM: 5.0%; ICD-10-CM: 15.0%; CPT: 9.0%). METEOR scores, BERTScores, and cui2vec similarity scores were also uniformly higher than in the full dataset (Table 2).

CODE GENERATION ERROR ANALYSIS

To further interrogate the observed code generation errors, we analyzed the incorrectly generated codes for Gemini Pro, Llama2-70b Chat, and the most recent model versions of GPT-3.5 and GPT-4. We found GPT-4 and GPT-3.5 Turbo had similarly high rates of valid codes, in contrast to Gemini Pro and Llama2-70b Chat, which had more difficulty (Table 3). Among the incorrect codes, GPT-4 and GPT-3.5 Turbo generated similar rates of nonexistent codes, followed by Gemini Pro and Llama2-70b Chat. Llama2-70b Chat had the highest rate of nonbillable codes (ICD-9-CM: 9.4%; ICD-10-CM: 35.1%), followed by Gemini Pro (ICD-9-CM: 11.1%; ICD-10-CM: 12.6%), GPT-4 (ICD-9-CM: 9.6%; ICD-10-CM: 12.7%), and GPT-3.5 Turbo (ICD-9-CM: 6.4%; ICD-10-CM: 10.4%). GPT-3.5 Turbo had the highest rate of billable codes for both ICD code systems, followed closely by GPT-4. Gemini Pro and Llama2-70b Chat had much lower rates of generated billable codes. Surprisingly, all models had a low rate (<75%) of correct length code (Table 3) for the code systems with variable code lengths (ICD-9-CM and ICD-10-CM). Overall digit-level matches were also low (<70%) across all models (Table 3). As a measure of code specificity, we determined the longest correct sequence and observed that the most frequent length was three digits for most scenarios (Table 3).

We observed the frequency of code repetition in the incorrectly generated codes (Table 3). We found GPT-4 had the lowest rate of code repetition and, for all models, ICD-9-CM codes repeated the least. Among incorrect codes, GPT-4 notably repeated ICD-10-CM codes a mean of 3.7 times in contrast to GPT-3.5 Turbo, which repeated codes a mean of 93.6 times. Table S4 shows the top five repeated codes for each model and code system. Interestingly, the most repeated CPT code for nearly all models was 84120, which corresponds to "Porphyrins, urine; quantitation and fractionation." We were unable to identify any pattern that explained which codes were repeated.

Automated textual similarity scores (METEOR score, BERTScore, and cui2vec cosine similarity) in the error analysis showed a similar pattern to what was observed in the overall dataset (Table 3). GPT-4 consistently scored highest and Llama2-70b Chat consistently scored lowest, with Gemini Pro and GPT-3.5 Turbo between these two extremes. Manually assessed CodeSTS scores were used to analyze conceptual similarity (Table 4). GPT-3.5 Turbo (Nov) had the highest rate of correctly generalized code matches (ICD-9-CM: 29.9%; ICD-10-CM: 18.5%), followed by GPT-4 (Nov) (ICD-9-CM: 18.6%; ICD-10-CM: 13.0%), Gemini Pro (ICD-9-CM: 9.2%; ICD-10-CM: 5.6%), and Llama2-70b Chat (ICD-9-CM: 1.6%; ICD-10-CM: 7.5%). We observed that in most instances a CodeSTS score of 1 was most frequent for the ICD code systems and a score of 0 for the CPT code system (Figure S1). In general, the GPT-4 models had the highest propensity for higher CodeSTS scores, which is consistent with other similarly high measures of conceptual similarity for these models. A full error analysis for the manually reviewed codes that were not exact matches is shown in Table S5.

EHR code frequency, shorter codes, and shorter code descriptions were generally associated with a higher rate of exact matches for all models and code systems ($P<0.05$). The exceptions were ICD-9-CM code length for Llama2-70b Chat and CPT description length for GPT-3.5 and GPT-4. Log-transformed EHR code frequency had the strongest correlation with ICD-9-CM and CPT code exact match accuracy for all models. Conversely, code length had the strongest correlation with ICD-10-CM code exact match accuracy. The distribution of exact match counts and rates for these code characteristics are illustrated in Figures S2 to S4.

Discussion

Our study evaluated the medical code–generating performance of GPT-3.5, GPT-4, Gemini Pro, and Llama2-70b Chat. We found that no model had an exact match rate for generated ICD-9-CM, ICD-10-CM, and CPT codes above 50%, rendering these models unsuitable in their base form. GPT-4 performed the best in terms of exact and equivalent match rates and multiple measures of conceptual similarity. GPT-3.5 was the next best model, followed by Gemini-Pro. Llama2-70b Chat performed the worst by a large margin, with exact match rates under 5%. All models generated CPT and ICD-9-CM codes more accurately than ICD-10-CM codes.

Table 2. ICD-9-CM, ICD-10-CM, and CPT Code Generation Overall Performance Metrics, Manually Reviewed Code Subset.*

Coding System	Metric	GPT-3.5 Turbo (March)†	GPT-3.5 Turbo (June)†	GPT-3.5 Turbo (Nov)†	GPT-4 (March)†	GPT-4 (June)†	GPT-4 (Nov)†	Gemini Pro†	Llama2-70b Chat†
ICD-9-CM (n=200)	Exact match, % (95% CI)	60.5% (53.5%–67.0%)	62.5% (55.5%–69.0%)	66.5% (60.0%–73.0%)	78.5% (72.5%–84.0%)	78.0% (72.0%–83.5%)	78.5% (72.5%–84.0%)	34.5% (28.0%–41.0%)	4.5% (2.0%–7.5%)
	Equivalent match, % (95% CI)	2.0% (0.5%–4.0%)	1.0% (0.0%–2.5%)	1.0% (0.0%–2.5%)	1.0% (0.0%–2.5%)	1.0% (0.0%–2.5%)	1.5% (0.0%–3.5%)	3.0% (1.0%–5.5%)	0.5% (0.0%–1.5%)
	CodeSTS score, mean (95% CI)	3.8 (3.6–4.1)	3.8 (3.6–4.0)	3.9 (3.7–4.2)	4.3 (4.0–4.5)	4.3 (4.1–4.5)	4.3 (4.1–4.5)	2.6 (2.3–2.8)	0.6 (0.5–0.8)
	cui2vec cosine similarity, mean (95% CI)	0.864 (0.831–0.896)	0.882 (0.852–0.912)	0.896 (0.866–0.924)	0.950 (0.927–0.969)	0.949 (0.930–0.966)	0.943 (0.918–0.965)	0.783 (0.743–0.819)	0.529 (0.469–0.591)
	METEOR score, mean (95% CI)	0.683 (0.626–0.740)	0.709 (0.652–0.765)	0.732 (0.677–0.785)	0.847 (0.800–0.891)	0.835 (0.786–0.880)	0.836 (0.789–0.882)	0.486 (0.423–0.549)	0.154 (0.107–0.208)
	BERTScore, mean (95% CI)	0.917 (0.901–0.933)	0.926 (0.910–0.940)	0.929 (0.914–0.945)	0.963 (0.951–0.974)	0.957 (0.945–0.969)	0.958 (0.946–0.970)	0.870 (0.851–0.888)	0.759 (0.742–0.776)
ICD-10-CM (n=200)	Exact match, % (95% CI)	56.5% (49.5%–63.5%)	60.0% (53.0%–67.0%)	59.0% (52.0%–66.0%)	73.5% (67.5%–79.5%)	71.5% (65.0%–77.5%)	75.5% (69.5%–81.5%)	27.0% (21.0%–33.5%)	13.0% (8.5%–18.0%)
	Equivalent match, % (95% CI)	2.5% (0.5%–5.0%)	1.5% (0.0%–3.5%)	2.0% (0.5%–4.0%)	1.0% (0.0%–2.5%)	0.5% (0.0%–1.5%)	2.5% (0.5%–5.0%)	0.5% (0.0%–1.5%)	2.0% (0.5%–4.0%)
	CodeSTS score, mean (95% CI)	3.6 (3.4–3.9)	3.8 (3.5–4.0)	3.7 (3.4–3.9)	4.1 (3.9–4.3)	4.1 (3.9–4.3)	4.3 (4.1–4.5)	2.1 (1.8–2.4)	1.7 (1.4–1.9)
	cui2vec cosine similarity, mean (95% CI)	0.879 (0.843–0.912)	0.873 (0.834–0.909)	0.860 (0.821–0.898)	0.916 (0.882–0.948)	0.913 (0.877–0.945)	0.938 (0.910–0.963)	0.730 (0.669–0.788)	0.486 (0.417–0.552)
	METEOR score, mean (95% CI)	0.723 (0.668–0.776)	0.719 (0.664–0.773)	0.702 (0.644–0.757)	0.817 (0.768–0.863)	0.807 (0.758–0.855)	0.855 (0.812–0.896)	0.500 (0.434–0.567)	0.283 (0.232–0.336)
	BERTScore, mean (95% CI)	0.938 (0.924–0.951)	0.934 (0.920–0.948)	0.931 (0.917–0.945)	0.958 (0.945–0.969)	0.959 (0.947–0.970)	0.970 (0.960–0.979)	0.886 (0.867–0.903)	0.814 (0.798–0.831)
CPT (n=200)	Exact match, % (95% CI)	47.5% (40.5%–54.5%)	42.0% (35.0%–49.0%)	63.0% (56.5%–69.5%)	69.0% (62.5%–75.5%)	70.0% (63.5%–76.0%)	83.5% (78.0%–88.5%)	28.0% (22.0%–34.5%)	9.0% (5.5%–13.0%)
	Equivalent match, % (95% CI)	0.0% (0.0%–0.0%)	0.0% (0.0%–0.0%)	0.0% (0.0%–0.0%)	0.0% (0.0%–0.0%)	0.0% (0.0%–0.0%)	0.0% (0.0%–0.0%)	0.0% (0.0%–0.0%)	0.0% (0.0%–0.0%)
	CodeSTS score, mean (95% CI)	2.7 (2.4–3.1)	2.5 (2.1–2.8)	3.6 (3.3–3.9)	3.8 (3.6–4.1)	3.8 (3.5–4.1)	4.5 (4.3–4.6)	2.1 (1.8–2.4)	0.9 (0.7–1.1)
	METEOR score, mean (95% CI)	0.598 (0.539–0.656)	0.536 (0.477–0.596)	0.742 (0.692–0.792)	0.768 (0.716–0.819)	0.773 (0.718–0.823)	0.915 (0.883–0.944)	0.467 (0.412–0.526)	0.265 (0.216–0.317)
	BERTScore, mean (95% CI)	0.896 (0.878–0.913)	0.877 (0.859–0.895)	0.941 (0.926–0.954)	0.943 (0.929–0.957)	0.942 (0.927–0.956)	0.986 (0.979–0.991)	0.867 (0.850–0.884)	0.798 (0.782–0.817)

* CI indicates confidence interval; CPT, Current Procedural Terminology; ICD-9-CM, International Classification of Diseases, 9th edition, Clinical Modification; ICD-10-CM, International Classification of Diseases, 10th edition, Clinical Modification; and Nov, November.

† Application programming interface accessed between December 26 and 27, 2023.

Despite struggling with exact code generation, the models often generated codes that were correct or at least conceptually similar to the correct codes. To complement the automated similarity scores, we created and applied the CodeSTS metric. We observed a moderate degree of interrater reliability between the two CodeSTS scorers. There was poor concordance between CodeSTS and the automated similarity scores, which we attribute to differences between each metric's definition of similarity. Within our manually reviewed subset, we found that GPT-3.5 had the greatest tendency to generate correct but generalized codes and GPT-4 had the greatest tendency to generate equivalent codes. Across the entire dataset, GPT-4 had the lowest rate of fabricated codes and GPT-3.5 had the lowest rate of nonbillable code generation. For GPT-4, GPT-3.5, and Gemini Pro, code generation errors were most likely after the first three generated digits. The CodeSTS score distributions confirm this observation, because most nonexact generated codes correspond with scores of 1 to 3, which all have some degree of relationship with the original code. The base LLMs can therefore parse the general descriptive nature of medical codes. Still, they often cannot attain adequate precision and resort to overgeneralization or fabricated specificity, which is unacceptable for clinical use cases.[23]

EHR code frequency had the largest impact on code-generated performance, likely explaining the differences in overall performance between the full dataset and manual subset. This relationship is likely due to the frequency of these codes appearing in the training data. Similarly, we noted that the models repeatedly generated the same codes for multiple descriptions, suggesting a tendency for specific codes or, potentially, gaps in the training data. We could not discern any pattern in the repeated codes. In general, code repetition was most frequent for CPT codes. Performance across code systems is most likely related to the frequency of each code and its description in the LLM training data.[24] The error patterns we observed suggest that the LLMs do not have a complete internal representation of medical coding rules. This is consistent with prior work showing that LLMs have difficulty performing multistep logic without support.[25] This is also consistent with our observations about model behavior while developing our prompt strategy. We found that, for particular code descriptions, no amount of prompt engineering could coerce the models to generate the correct code.

Our results are consistent with a prior experiment showing that smaller, fine-tuned Spark NLP models (76% capture rate) outperformed GPT-4 (58%) and GPT-3.5 (40%) when extracting ICD-10-CM codes from a limited set of clinical text.[11] We similarly found a high rate of incorrectly generated ICD codes. However, our approach differs in several key aspects. By simplifying the model task to match codes to their descriptions (code querying), we could scale our benchmarking to over 27,000 ICD and CPT codes. We also focused our study on explaining why more advanced LLMs would perform worse at extracting codes from clinical text. We suspected that because general-purpose LLMs struggle with tasks requiring character-level comprehension, such as arithmetic or word spelling,[26-28] they would similarly struggle with the fundamental task of matching alphanumeric medical codes to their official descriptions. This task isolates the model's understanding of the medical codes from higher-level language tasks like understanding clinical concepts.

Our study reaffirms the limitations of LLM tokenization. LLMs are trained on and generate text in short segments, achieved by splitting the source text into basic linguistic units known as "tokens." However, when tokenization is applied to nonlanguage text such as medical codes, it clusters the characters without regard for the coding system's intrinsic structure and obscures that information from the model.[29,30] The limitations of tokenization may be overcome with additional LLM fine-tuning or linkage with programmed "tools."[31-34]

Our study has several limitations. We did not evaluate strategies known to improve LLM performance, including advanced prompt engineering, tool use, retrieval augmented generation, or model fine-tuning. We also did not evaluate code generation performance in the context of real-world clinical narratives or notes from the EHR. Despite this, our validation approach targets a key bottleneck in performance and provides scalable and reproducible open-source evaluation metrics that can spur accelerated development of LLM-based medical code extraction tools. Since our study was conducted, new and updated models have been released that may perform better on medical code querying.

Our evaluation of LLM proficiency in generating codes from the ICD-9-CM, ICD-10-CM, and CPT systems

Table 3. ICD-9-CM, ICD-10-CM, and CPT Code Generation Error Analysis of Current Model Versions, Full Code Set.*

Coding System	Metric	GPT-3.5 Turbo (Nov)†	GPT-4 (Nov)†	Gemini Pro†	Llama2-70b Chat†
ICD-9-CM (n=7697)	Incorrect codes, n (% of total)	5467 (71.0%)	4149 (53.9%)	6869 (89.2%)	7601 (98.8%)
	Valid code, % (95% CI)	96.1% (95.6%–96.6%)	97.1% (96.6%–97.5%)	88.9% (88.1%–89.6%)	54.1% (53.0%–55.2%)
	Billable code, % (95% CI)	89.7% (88.9%–90.5%)	87.5% (86.5%–88.5%)	69.8% (68.7%–70.9%)	44.7% (43.6%–45.8%)
	Nonbillable code, % (95% CI)	6.4% (5.8%–7.0%)	9.6% (8.7%–10.5%)	19.1% (18.2%–20.0%)	9.4% (8.8%–10.1%)
	Fabricated code, % (95% CI)	3.9% (3.4%–4.4%)	2.9% (2.4%–3.5%)	11.1% (10.4%–11.8%)	45.9% (44.8%–47.0%)
	Code generation frequency, mean (95% CI)	4.9 (4.7–5.0)	3.0 (3.0–3.1)	6.5 (6.3–6.6)	17.5 (16.9–18.1)
	Matched length, % (95% CI)	71.8% (70.6%–73.0%)	73.9% (72.5%–75.2%)	62.7% (61.5%–63.8%)	58.1% (57.0%–59.2%)
	Matched digits, % (95% CI)	56.3% (55.6%–57.0%)	63.3% (62.6%–64.0%)	53.2% (52.6%–53.8%)	30.8% (30.2%–31.4%)
	Longest sequence of correct digits matched, % (95% CI)	0: 18.7% (17.7%–19.7%)	0: 10.0% (9.1%–10.9%)	0: 17.1% (16.2%–18.0%)	0: 42.2% (41.1%–43.3%)
		1: 5.9% (5.3%–6.6%)	1: 4.9% (4.3%–5.6%)	1: 9.9% (9.2%–10.6%)	1: 21.2% (20.3%–22.2%)
		2: 12.7% (11.8%–13.6%)	2: 12.5% (11.5%–13.5%)	2: 20.8% (19.8%–21.7%)	2: 21.2% (20.3%–22.1%)
		3: 42.3% (41.0%–43.6%)	3: 42.0% (40.5%–43.6%)	3: 36.5% (35.3%–37.6%)	3: 12.3% (11.6%–13.1%)
		4: 20.4% (19.3%–21.4%)	4: 30.6% (29.3%–32.0%)	4: 15.8% (14.9%–16.6%)	4: 3.1% (2.7%–3.5%)
	cui2vec cosine similarity, mean (95% CI)	0.660 (0.654–0.667)	0.697 (0.690–0.703)	0.590 (0.584–0.596)	0.402 (0.393–0.411)
	METEOR score, mean (95% CI)	0.198 (0.193–0.202)	0.235 (0.230–0.240)	0.153 (0.150–0.157)	0.079 (0.075–0.083)
	BERTScore, mean (95% CI)	0.805 (0.803–0.807)	0.824 (0.822–0.827)	0.786 (0.784–0.788)	0.743 (0.741–0.745)
ICD-10-CM (n=15,950)	Incorrect codes, n (% of total)	13,025 (81.7%)	10,492 (65.8%)	15,170 (95.1%)	15,693 (98.4%)
	Valid code, % (95% CI)	82.7 (82.0%–83.3%)	81.5 (80.7%–82.2%)	62.6 (61.8%–63.4%)	69.7 (69.0%–70.4%)
	Billable code, % (95% CI)	72.2% (71.4%–73.0%)	68.8% (67.9%–69.7%)	50.0% (49.2%–50.8%)	34.6% (33.9%–35.4%)
	Nonbillable code, % (95% CI)	10.4% (9.9%–11.0%)	12.7% (12.0%–13.3%)	12.6% (12.1%–13.2%)	35.1% (34.3%–35.8%)
	Fabricated code, % (95% CI)	17.3% (16.7%–18.0%)	18.5% (17.8%–19.2%)	37.4% (36.6%–38.2%)	30.3% (29.6%–31.0%)
	Code generation frequency, mean (95% CI)	93.6 (88.6–98.7)	3.7 (3.7–3.8)	46.2 (44.0–48.4)	63.1 (60.4–65.9)
	Matched length, % (95% CI)	57.4% (56.6%–58.3%)	64.7% (63.8%–65.7%)	58.9% (58.1%–59.7%)	31.3% (30.6%–32.1%)
	Matched digits, % (95% CI)	57.0% (56.6%–57.4%)	67.6% (67.2%–68.0%)	51.6% (51.3%–52.0%)	37.5% (37.1%–37.8%)
	Longest sequence of correct digits matched, % (95% CI)	0: 13.7% (13.1%–14.3%)	0: 5.0% (4.6%–5.4%)	0: 12.0% (11.5%–12.5%)	0: 20.7% (20.1%–21.4%)

(continued)

Table 3. (cont.)

Coding System	Metric	GPT-3.5 Turbo (Nov)†	GPT-4 (Nov)†	Gemini Pro†	Llama2-70b Chat†
		1: 7.1% (6.7%–7.5%)	1: 2.9% (2.6%–3.2%)	1: 14.3% (13.7%–14.8%)	1: 26.6% (25.9%–27.2%)
		2: 9.1% (8.6%–9.6%)	2: 5.6% (5.2%–6.1%)	2: 18.6% (18.0%–19.2%)	2: 24.1% (23.5%–24.8%)
		3: 44.8% (43.9%–45.6%)	3: 42.1% (41.1%–43.0%)	3: 40.6% (39.9%–41.4%)	3: 23.6% (22.9%–24.2%)
		4: 18.5% (17.8%–19.2%)	4: 29.9% (29.0%–30.7%)	4: 12.4% (11.9%–12.9%)	4: 4.6% (4.3%–4.9%)
		5: 5.8% (5.4%–6.2%)	5: 11.6% (11.0%–12.2%)	5: 2.0% (1.7%–2.2%)	5: 0.4% (0.3%–0.4%)
		6: 1.0% (0.8%–1.2%)	6: 2.9% (2.6%–3.2%)	6: 0.1% (0.1%–0.2%)	6: 0.0% (0.0%–0.1%)
	cui2vec cosine similarity, mean (95% CI)	0.417 (0.411–0.423)	0.512 (0.506–0.518)	0.349 (0.342–0.355)	0.255 (0.249–0.261)
	METEOR score, mean (95% CI)	0.237 (0.234–0.240)	0.314 (0.310–0.317)	0.188 (0.185–0.191)	0.108 (0.106–0.110)
	BERTScore, mean (95% CI)	0.830 (0.828–0.831)	0.866 (0.864–0.868)	0.810 (0.808–0.811)	0.769 (0.768–0.770)
CPT (n=3673)	Incorrect codes, n (% of total)	2502 (68.1%)	1843 (50.2%)	3225 (88.6%)	3579 (97.4%)
	Valid code, % (95% CI)	94.0% (93.0%–94.9%)	93.9% (92.8%–95.0%)	84.1% (82.8%–85.3%)	54.8% (53.1%–56.4%)
	Fabricated code, % (95% CI)	6.0% (5.1%–7.0%)	6.1% (5.0%–7.2%)	15.9% (14.7%–17.2%)	45.2% (43.6%–46.9%)
	Code Generation Frequency, mean (95% CI)	8.4 (7.5–9.3)	2.6 (2.5–2.7)	15.3 (14.0–16.7)	60.3 (56.3–64.4)
	Matched length, % (95% CI)	99.7% (99.4%–99.9%)	98.5% (98.0%–99.1%)	98.7% (98.3%–99.1%)	98.8% (98.4%–99.1%)
	Matched digits, % (95% CI)	59.5% (58.7%–60.4%)	63.3% (62.3%–64.2%)	53.7% (53.0%–54.5%)	40.8% (40.1%–41.6%)
	Longest sequence of correct digits matched, % (95% CI)	0: 4.2% (3.5%–5.0%)	0: 4.4% (3.5%–5.4%)	0: 7.5% (6.6%–8.4%)	0: 13.2% (12.1%–14.3%)
		1: 12.0% (10.7%–13.2%)	1: 6.5% (5.4%–7.7%)	1: 15.5% (14.3%–16.8%)	1: 32.6% (31.0%–34.1%)
		2: 18.5% (17.0%–20.0%)	2: 16.0% (14.3%–17.6%)	2: 27.7% (26.2%–29.2%)	2: 29.6% (28.1%–31.1%)
		3: 34.8% (32.9%–36.6%)	3: 36.0% (33.8%–38.2%)	3: 30.1% (28.5%–31.7%)	3: 18.2% (16.9%–19.4%)
		4: 30.5% (28.7%–32.4%)	4: 37.2% (34.9%–39.4%)	4: 19.2% (17.8%–20.6%)	4: 6.5% (5.8%–7.3%)
	METEOR score, mean (95% CI)	0.243 (0.236–0.251)	0.289 (0.281–0.298)	0.187 (0.182–0.193)	0.143 (0.137–0.149)
	BERTScore, mean (95% CI)	0.817 (0.813–0.821)	0.838 (0.834–0.842)	0.788 (0.785–0.791)	0.759 (0.756–0.762)

* CI denotes confidence interval; CPT, Current Procedural Terminology; ICD-9-CM, International Classification of Diseases, 9th edition, Clinical Modification; ICD-10-CM, International Classification of Diseases, 10th edition, Clinical Modification; and Nov, November.

† Application programming interface accessed between December 26 and 27, 2023.

suggests that base LLMs alone are poorly suited for medical code mapping tasks. Although the models can approximate the meaning of many codes, they also display an unacceptable lack of precision and a high propensity for falsifying codes. This has significant implications for billing, clinical decision-making, quality improvement, research, and health policy. Whereas we found that current base LLMs struggle with simple code queries, there is an opportunity to mitigate this with fine-tuning, tool use, or retrieval augmented generation. Finally, we provide a systematic and automated evaluation approach for medical code generation that can spur the development of medical code extraction tools.

Table 4. ICD-9-CM, ICD-10-CM, and CPT Code Generation Error Analysis of Current Model Versions, Manually Reviewed Code Subset.*					
Coding System	**Metric**	**GPT-3.5 Turbo (Nov)†**	**GPT-4 (Nov)†**	**Gemini Pro†**	**Llama2-70b Chat†**
ICD-9-CM (n=200)	Incorrect codes, n (% of total)	67 (33.5%)	43 (21.5%)	131 (65.5%)	191 (95.5%)
	Valid code, % (95% CI)	95.5% (89.6%–100.0%)	93.0% (83.7%–100.0%)	82.4% (75.6%–88.5%)	55.0% (48.2%–61.8%)
	Billable code, % (95% CI)	91.0% (83.6%–97.0%)	83.7% (72.1%–93.0%)	62.6% (54.2%–71.0%)	44.5% (37.7%–51.3%)
	Equivalent match, % (95% CI)	3.0% (0.0%–7.5%)	7.0% (0.0%–16.3%)	4.6% (1.5%–8.4%)	0.5% (0.0%–1.6%)
	Generalized match, % (95% CI)	29.9% (19.4%–40.3%)	18.6% (7.0%–30.2%)	9.2% (4.6%–14.5%)	1.6% (0.0%–3.7%)
	Nonbillable code, % (95% CI)	4.5% (0.0%–10.4%)	9.3% (2.3%–18.6%)	19.8% (13.0%–26.7%)	10.5% (6.3%–15.2%)
	Fabricated code, % (95% CI)	4.5% (0.0%–10.4%)	7.0% (0.0%–16.3%)	17.6% (11.5%–24.4%)	45.0% (38.2%–52.4%)
	CodeSTS score, mean (95% CI)	1.9 (1.6–2.1)	1.9 (1.5–2.3)	1.3 (1.1–1.5)	0.4 (0.3–0.5)
ICD-10-CM (n=200)	Incorrect codes, n (% of total)	81 (40.5%)	46 (23%)	144 (72%)	173 (86.5%)
	Valid code, % (95% CI)	87.7% (80.2%–93.8%)	84.8% (73.9%–93.5%)	63.9% (56.2%–71.5%)	79.2% (72.8%–85.0%)
	Billable code, % (95% CI)	76.5% (66.7%–85.2%)	65.2% (52.2%–78.3%)	47.9% (39.6%–56.2%)	49.1% (41.6%–56.6%)
	Equivalent match, % (95% CI)	4.9% (1.2%–9.9%)	10.9% (2.2%–19.6%)	0.7% (0.0%–2.1%)	2.3% (0.6%–4.6%)
	Generalized match, % (95% CI)	18.5% (9.9%–27.2%)	13.0% (4.3%–23.9%)	5.6% (2.1%–9.7%)	7.5% (4.0%–11.6%)
	Nonbillable code, % (95% CI)	11.1% (4.9%–18.5%)	19.6% (8.7%–30.4%)	16.0% (10.4%–22.2%)	30.1% (23.1%–37.0%)
	Fabricated code, % (95% CI)	12.3% (6.2%–19.8%)	15.2% (6.5%–26.1%)	36.1% (28.5%–44.4%)	20.8% (15.0%–26.6%)
	CodeSTS score, mean (95% CI)	1.7 (1.5–2.0)	1.8 (1.4–2.2)	0.9 (0.8–1.1)	1.1 (1.0–1.3)
CPT (n=200)	Incorrect codes, n (% of total)	94.6% (89.2%–98.6%)	84.8% (72.7%–97.0%)	86.1% (80.6%–91.7%)	74.2% (67.6%–80.2%)
	Valid code, % (95% CI)	0.0% (0.0%–0.0%)	0.0% (0.0%–0.0%)	0.0% (0.0%–0.0%)	0.0% (0.0%–0.0%)
	Equivalent match, % of (95% CI)	6.8% (1.4%–13.5%)	15.2% (3.0%–27.3%)	10.4% (5.6%–16.0%)	2.7% (0.5%–5.5%)
	Fabricated code, % of (95% CI)	5.4% (1.4%–10.8%)	15.2% (3.0%–27.3%)	13.9% (8.3%–19.4%)	25.8% (19.8%–32.4%)
	CodeSTS score, mean (95% CI)	1.2 (1.0–1.4)	1.8 (1.4–2.1)	1.0 (0.8–1.2)	(0.3–0.6)

* CI denotes confidence interval; CPT, Current Procedural Terminology; ICD-9-CM, International Classification of Diseases, 9th edition, Clinical Modification; ICD-10-CM, International Classification of Diseases, 10th edition, Clinical Modification; and Nov, November.

† Application programming interface accessed between December 26 and 27, 2023.

Disclosures

Supported by the AGA Research Foundation's 2023 AGA-Amgen Fellowship-to-Faculty Transition Award AGA2023-32-06 and a National Institutes of Health UL1TR004419 award.

Author disclosures and other supplementary materials are available at ai.nejm.org.

The extracted medical codes, code generation pipeline, and our analyses are all freely available at https://github.com/Nadkarni-Lab/LLM_CodeQuery. Code frequencies are available upon request.

Author Affiliations

[1] Division of Data-Driven and Digital Medicine (D3M), Icahn School of Medicine at Mount Sinai, New York

[2] The Charles Bronfman Institute of Personalized Medicine, Icahn School of Medicine at Mount Sinai, New York

[3] Henry D. Janowitz Division of Gastroenterology, Icahn School of Medicine at Mount Sinai, New York

[4] Central Management, Sheba Medical Centre, Ramat Gan, Israel

[5] Faculty of Medicine, Tel Aviv University, Tel Aviv, Israel

[6] Department of Diagnostic Imaging, Sheba Medical Center, Ramat Gan, Israel

[7] Mount Sinai Health System, New York

[8] ARC Center for Digital Innovation, Sheba Medical Center, Ramat Gan, Israel

References

1. World Health Organization. History of the development of the ICD. (https://cdn.who.int/media/docs/default-source/classification/icd/historyoficd.pdf).
2. World Health Organization. Importance of ICD. 2024 (https://www.who.int/standards/classifications/frequently-asked-questions/importance-of-icd).
3. Wood PH. Applications of the International Classification of Diseases. World Health Stat Q 1990;43:263-268.
4. American Medical Association. CPT® overview and code approval. 2023 (https://www.ama-assn.org/practice-management/cpt/cpt-overview-and-code-approval).
5. Dong H, Falis M, Whiteley W, et al. Automated clinical coding: what, why, and where we are? NPJ Digit Med 2022;5:159. DOI: 10.1038/s41746-022-00705-7.
6. Zhao WX, Zhou K, Li J, et al. A survey of large language models. November 24, 2023 (https://arxiv.org/abs/2303.18223). Preprint.
7. Thirunavukarasu AJ, Ting DSJ, Elangovan K, Gutierrez L, Tan TF, Ting DSW. Large language models in medicine. Nat Med 2023;29:1930-1940. DOI: 10.1038/s41591-023-02448-8.
8. Lee P, Bubeck S, Petro J. Benefits, limits, and risks of GPT-4 as an AI chatbot for medicine. N Engl J Med 2023;388:1233-1239. DOI: 10.1056/NEJMsr2214184.
9. Nori H, King N, McKinney SM, Carignan D, Horvitz E. Capabilities of GPT-4 on medical challenge problems. March 20, 2023 (https://arxiv.org/abs/2303.13375). Preprint.
10. Singhal K, Azizi S, Tu T, et al. Large language models encode clinical knowledge [published correction appears in Nature 2023;620:E19]. Nature 2023;620:172-180. DOI: 10.1038/s41586-023-06291-2.
11. Kocaman V. Comparing Spark NLP for healthcare and ChatGPT in extracting ICD10-CM codes from clinical notes. 2023 (https://www.johnsnowlabs.com/comparing-spark-nlp-for-healthcare-and-chatgpt-in-extracting-icd10-cm-codes-from-clinical-notes/).
12. Bodenreider O. The Unified Medical Language System (UMLS): integrating biomedical terminology. Nucleic Acids Res 2004;32:D267-D270. DOI: 10.1093/nar/gkh061.
13. Open AI. GPT-4 technical report. March 15, 2023 (https://arxiv.org/abs/2303.08774). Preprint.
14. Ouyang L, Wu J, Jiang X, et al. Training language models to follow instructions with human feedback. Adv Neural Inf Process Syst 2022;35:27730-27744.
15. Anil R, Borgeaud S, Wu Y, et al. Gemini: a family of highly capable multimodal models. December 19, 2023 (https://arxiv.org/abs/2312.11805). Preprint.
16. Touvron H, Martin L, Stone K, et al. Llama 2: open foundation and fine-tuned chat models. July 18, 2023 (https://arxiv.org/search/cs?searchtype=author&query=Kerkez,+V). Preprint.
17. White J, Fu Q, Hays S, et al. A prompt pattern catalog to enhance prompt engineering with ChatGPT. February 21, 2023 (https://arxiv.org/abs/2302.11382). Preprint.
18. Wang Y, Afzal N, Fu S, et al. MedSTS: a resource for clinical semantic textual similarity. Lang Resour Eval 2020;54:57-72. DOI: 10.1007/s10579-018-9431-1.
19. Denkowski M, Lavie A. Meteor Universal: language specific translation evaluation for any target language. In: Bojar O, Buck C, Federmann C, et al., eds. Proceedings of the Ninth Workshop on Statistical Machine Translation. Baltimore, MD: Association for Computational Linguistics, 2014:376-380. DOI: 10.3115/v1/W14-3348.
20. Zhang T, Kishore V, Wu F, Weinberger KQ, Artzi Y. BERTscore: evaluating text generation with BERT. 2019 (https://arxiv.org/abs/1904.09675). Preprint.
21. Beam AL, Kompa B, Schmaltz A, et al. Clinical concept embeddings learned from massive sources of multimodal medical data. Pac Symp Biocomput 2020;25:295-306.
22. Centers for Medicare & Medicaid Services. ICD code lists. 2022 (https://www.cms.gov/medicare/coordination-benefits-recovery/overview/icd-code-lists).
23. McKenna N, Li T, Cheng L, Hosseini MJ, Johnson M, Steedman M. Sources of hallucination by large language models on inference tasks. October 22, 2023 (https://arxiv.org/abs/2305.14552). Preprint.

24. Razeghi Y, Logan RL, Gardner M, Singh S. Impact of pretraining term frequencies on few-shot reasoning. May 24, 2022 (https://arxiv.org/abs/2202.07206). Preprint.

25. Huang J, Chang KC-C. Towards reasoning in large language models: a survey. May 26, 2023 (https://arxiv.org/abs/2212.10403). Preprint.

26. Yuan Z, Yuan H, Tan C, Wang W, Huang S. How well do large language models perform in arithmetic tasks? March 16, 2023 (https://arxiv.org/abs/2304.02015). Preprint.

27. Kim J, Hong G, Kim K, Kang J, Myaeng S-H. Have you seen that number? Investigating extrapolation in question answering models. 2021 (https://aclanthology.org/2021.emnlp-main.563/).

28. Nogueira R, Jiang Z, Lin J. Investigating the limitations of transformers with simple arithmetic tasks. April 12, 2021 (https://arxiv.org/abs/2102.13019). Preprint.

29. OpenAI. Tokenizer. (https://platform.openai.com/tokenizer).

30. Brown TB, Mann B, Ryder N, et al. Language models are few-shot learners. May 28, 2020 (https://arxiv.org/abs/2005.14165). Preprint.

31. Yang K, Liu J, Wu J, et al. If LLM is the wizard, then code is the wand: a survey on how code empowers large language models to serve as intelligent agents. January 8, 2024 (https://arxiv.org/abs/2401.00812). Preprint.

32. Peng B, Galley M, He P, et al. Check your facts and try again: improving large language models with external knowledge and automated feedback. March 8, 2023 (https://arxiv.org/abs/2302.12813). Preprint.

33. Huang J, Gu SS, Hou L, et al. Large language models can self-improve. October 25, 2022 (https://arxiv.org/abs/2210.11610). Preprint.

34. Lu P, Peng B, Cheng H, et al. Chameleon: plug-and-play compositional reasoning with large language models. October 31, 2023 (https://arxiv.org/abs/2304.09842). Preprint.