ESC European Heart Journal - Digital Health (2026) 7, ztag031
https://doi.org/10.1093/ehjdh/ztag031

ORIGINAL ARTICLE

Artificial intelligence-based automated interpretation of images of electrocardiograms: development and multinational validation of ECG-GPT

Akshay Khunte [1,2,†], Veer Sangha [1,3,†], Evangelos K. Oikonomou [1], Lovedeep S. Dhingra [1], Arya Aminorroaya [1], Andreas Coppi [1,4], Sumukh Vasisht Shankar [1], Elijah Rockers [5], Bobak J. Mortazavi [4,6], Deepak L. Bhatt [7], Harlan M. Krumholz [1,4,8], Sadeer Al-Kindi [5], Girish N. Nadkarni [9,10,‡], Akhil Vaid [9,10,‡], and Rohan Khera [1,4,11,*‡]

[1]Section of Cardiovascular Medicine, Department of Internal Medicine, Yale School of Medicine, 195 Church St, 6th Floor, New Haven, CT 06510, USA; [2]NYU Grossman School of Medicine, New York, NY, USA; [3]Department of Engineering Science, Oxford University, Oxford, UK; [4]Center for Outcomes Research and Evaluation, Yale-New Haven Hospital, New Haven, CT, USA; [5]Center for Cardiovascular Computational & Precision Health, Houston Methodist DeBakey Heart & Vascular Center, Houston, TX, USA; [6]Department of Computer Science & Engineering, Texas A&M University, College Station, TX, USA; [7]Mount Sinai Fuster Heart Hospital, Icahn School of Medicine at Mount Sinai, New York, NY, USA; [8]Department of Health Policy and Management, Yale School of Public Health, New Haven, CT, USA; [9]Windreich Department of Artificial Intelligence and Human Health, Icahn School of Medicine at Mount Sinai, New York, NY 10029, USA; [10]The Hasso Plattner Institute for Digital Health at Mount Sinai, New York, NY 10029, USA; and [11]Section of Health Informatics, Department of Biostatistics, Yale School of Public Health, New Haven, CT, USA

Received 19 August 2025; revised 26 September 2025; accepted 16 November 2025; online publish-ahead-of-print 18 February 2026

Aims

Timely, accurate assessment of electrocardiograms (ECGs) is crucial for diagnosing, triaging, and managing patients. However, this often relies on expert interpretation, a major bottleneck in low-resource settings. We developed and validated ECG-GPT, a format-independent vision encoder–decoder model that generates expert-level interpretations from 12-lead ECG images.

Methods and results

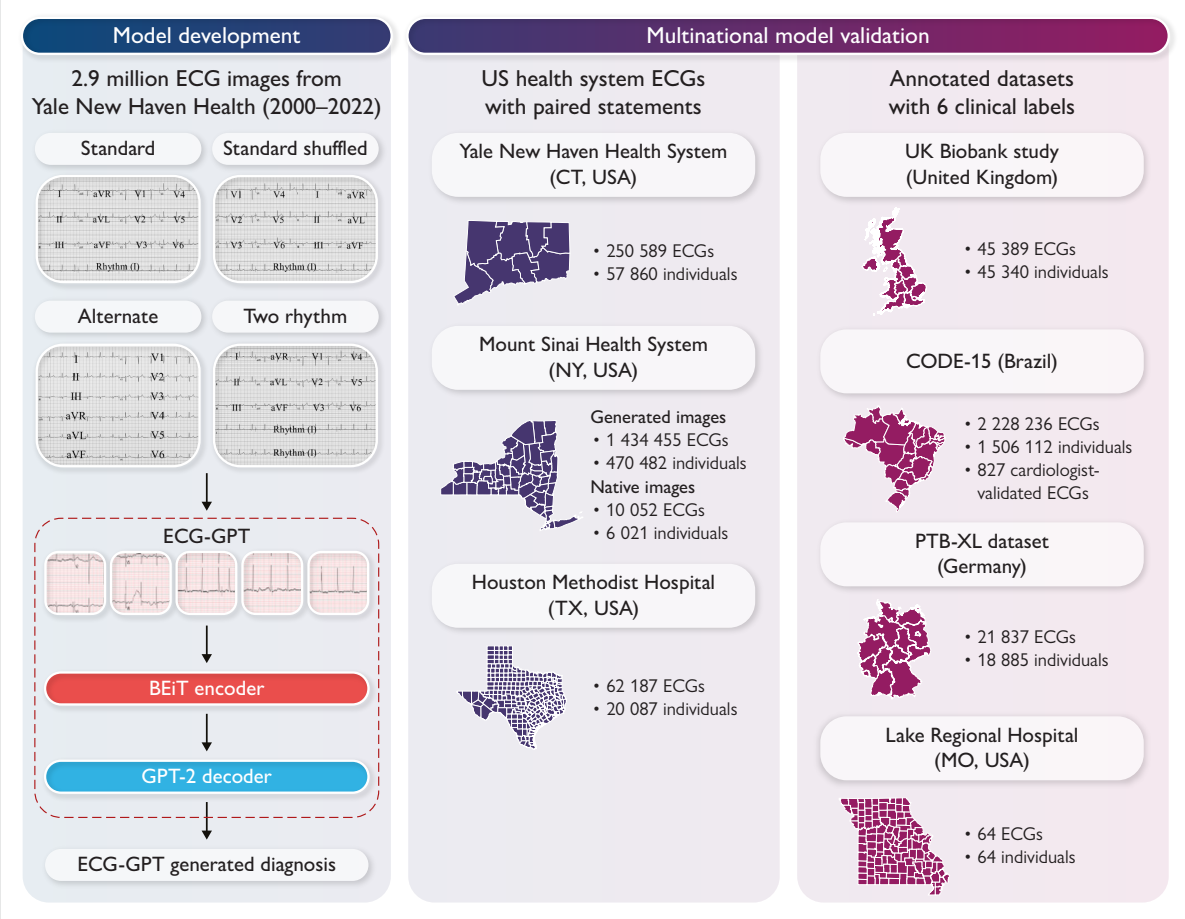
We developed ECG-GPT using 12-lead ECGs and their corresponding diagnosis statements performed at a large US health system between 2000 and 2022. Using structured clinical assessment, semantic similarity, and conventional metrics, we validated ECG-GPT across seven distinct health settings, including three large and diverse US health systems, ECGs from Minas Gerais, Brazil, the UK Biobank, the Germany-based PTB-XL dataset, and a community hospital in Missouri. In total, 2.9 million ECGs were used for model development, and 4.1 million ECGs for validation. The model performed well in clinical assessment across 26 extracted labels, with diagnostic accuracy ranging from 0.93 to 0.99. For rhythm abnormalities, including atrial fibrillation, sinus tachycardia, sinus bradycardia, premature atrial contractions, and premature ventricular contractions, AUROCs ranged from 0.80 to 0.95. For conduction abnormalities, including left bundle branch block, right bundle branch block, first degree atrioventricular block, left anterior fascicular block, and left posterior fascicular block, AUROCs ranged from 0.88 to 0.96. ECG-GPT identified the full context of

* Corresponding author. Tel: +1 203-785-4114, Email: rohan.khera@yale.edu, @rohan_khera
†Contributed equally as co-first authors
‡Contributed equally as co-senior authors

diagnosis statements with allied conditions with a median pairwise similarity of 0.90, significantly greater than baseline ($P < 0.001$). Results were comparable across external validation sites.

Conclusion We developed and validated a vision encoder-decoder model that generates expert-level interpretations from ECG images, a scalable strategy for accessible automated ECG analysis.

Structured Graphical Abstract



Keywords Artificial intelligence • Digital health • Electrocardiography

Introduction

Electrocardiography (ECG) is a widely available, first-line, non-invasive tool for diagnosing, triaging, and managing cardiovascular disease.[1] Traditional workflows often rely on computerized ECG interpretation algorithms to generate preliminary reads, which, despite limited accuracy,[2,3] provide diagnostic support and enable faster triage for high-risk conditions.[4] Such algorithms, however, are often proprietary, require raw signal data, and have substantial variability in accuracy,[4,5] with the potential to lead to patient harm.[6,7] This often makes computerized pre-reads inaccessible to clinicians in community health centers in rural areas of low- to middle-income countries, where fewer expert-level readers are available.[8,9] In these settings, management frequently relies on telecommunications with a limited number of experts available to provide such consultation.[10] Moreover, the accuracy of ECG interpretations by physicians is variable even after educational interventions, with less than 75% accuracy reported for even cardiologists.[11] Therefore, there is an unmet need for accessible, reliable, and accurate tools to provide expert-level ECG interpretations.

Though recent advances in deep learning enable accurate classification of specific ECG abnormalities,[12–15] they are generally limited to a select number of commonly encountered abnormalities, rather than the full range of abnormalities and modifying conditions. Additionally, while our prior work has demonstrated that diagnostic models can directly identify information from ECG images,[15,16] traditional signal-based models have limited scalability to the point-of-care and across low-resource

settings. The development of models exclusively for images is challenged by variations in the layout and labeling of the leads, the graphed background, and image quality.

In this study, we report the development of ECG-GPT (*Figure 1*), a novel vision-text transformer model capable of generating diagnostic reports from ECG images regardless of the layout, trained against the full breadth of expert-verified ECG interpretations across 2.9 million 12-lead ECGs collected over 21 years in a large US-based hospital system. ECG-GPT can be accessed as a web-based application that can receive ECG images across formats as the only input and generate diagnostic reports (demonstration hosted at https://www.cards-lab.org/ecg-gpt). We pursued broad multinational validation in 4.1 million ECGs across temporally and geographically distinct datasets, including three distinct US-based major referral hospital systems, a Brazil-based large telehealth network, a UK-based prospective cohort study, a publicly available ECG dataset from Germany, and a rural US-based community hospital.

Methods

The study was reviewed by the Yale Institutional Review Board, which approved the protocol and waived the need for informed consent as the study represents secondary analysis of existing data. Given the stipulations of the relevant institutional review boards, the datasets from Yale and the validation datasets from Mount Sinai, Houston Methodist Hospital, and Lake Regional Hospital are not publicly available. The external datasets from Brazil, UK Biobank, and PTB-XL are available directly from the respective groups and are outside the purview of the authors.

Data source for model development

Raw voltage data were collected for all 12-lead ECGs with corresponding diagnosis statements obtained at the Yale New Haven Health System (YNHHS) during 2000–2022. The subset of these ECGs with continuous recording across all 12 leads was then split at the patient level into training, validation, and test sets (85%, 5%, and 10%). For the test set, we restricted to ECGs with confirmed reports certified by a cardiologist. In the training and validation sets, we included abnormal, unconfirmed ECGs collected in the emergency setting, since these are interpreted at the point-of-care, and may not be listed as a confirmed read in the system. Finally, in the training set, a random set of ECGs with no marked abnormalities was down-sampled to match the observed prevalence of such ECGs in the original cohort (details in Supplementary material online, *Figure S1*).

Data pre-processing

To enable the generation of ECG images like those used in clinical settings, we implemented a pipeline to pre-process each signal before plotting (appendix p 2). Next, we used an approach we previously developed and validated to convert each signal waveform into multiple images using four lead layouts to account for different schemes of real-world ECGs, with this approach previously shown to generalize to formats not observed during development.[15] For model inference on native ECG images, we implemented a previously validated approach to standardize model inputs.[16] Finally, all diagnosis statements used for model development and validation were standardized using a rule-based approach to remove identifying information, expand abbreviations, and correct misspellings (see Supplementary material online, *Figure S2*).

Model development

We developed a custom Vision Encoder-Decoder model with over 239 million trainable parameters to generate statements directly from ECG images (*Figure 1*). This consisted of a BEiT encoder with a 384×384 pixel input size,[17] a Generative Pretrained Transformer-2 (GPT-2) transformer for the text decoder,[18] and a fine-tuned GPT-2 tokenizer (appendix pp 2–3). The model was trained on images plotted in randomly selected formats for 20 epochs at a learning rate of 5×10^{-5}.

External validation

We pursued validation on five generated ECG image datasets and three native ECG image datasets acquired outside the YNHHS. A detailed description of each external validation dataset is provided in the appendix (p 3). Briefly, these included A) one generated image dataset and one distinct native image dataset from Mount Sinai Health System (MSHS) in New York, NY, B) a native image dataset from Houston Methodist Hospital (HMH) in Houston, TX, C) a native image dataset from Lake Regional Hospital (LRH) in Osage Beach, MO, and D) images generated from four publicly available ECG signal datasets: the UK Biobank, the CODE15 dataset and a secondary cardiologist-validated dataset from Brazil, and the PTB-XL dataset from Germany. The ECGs from MSHS and HMH each had corresponding diagnosis statements, while the other datasets included labels for six rhythm and conduction disorders: atrial fibrillation (AF), sinus tachycardia (ST), sinus bradycardia (SB), left bundle branch block (LBBB), right bundle branch block (RBBB), and atrioventricular block (AVb).

Model evaluation

First, clinical labels for 26 conditions, spanning key rhythm and conduction disorders selected by two cardiologists, were extracted from each statement using a standardized string search approach (appendix p 4). Then, we evaluated ECG-GPT using three distinct strategies. These included (1) semantic similarity to assess the similarity between reference and model-generated statements and the model's ability to diagnose specific conditions within their clinical context. (2) The use of conventional natural language generation metrics, including ROUGE, BLEU, METEOR, and CIDEr scores,[19–22] to quantify the syntactic similarity of reference and generated statements. (3) A structured label assessment to evaluate the model's accuracy for the extracted labels using multiple metrics, including the area under the receiver operating characteristic (AUROC), area under precision–recall curve (AUPRC), accuracy, sensitivity, specificity, F1 score, positive predictive value (PPV), and negative predictive value (NPV).

Statistical analysis

Summary statistics are presented as counts and percentages for categorical elements and median and interquartile range (IQR) for continuous elements. A paired *t*-test was used to compute the probability of overlap between the pairwise and baseline cosine similarity of reference and model-generated statements. All analyses were performed using Python 3.11.3, and the significance level was set at an alpha of 0.05.

Results

Study population

We developed ECG-GPT in a set of 2,888,384 12-lead ECG recordings with accompanying cardiologist-confirmed diagnosis statements performed on 601 616 unique patients at the YNHHS between 2000 and 2022. These ECGs reflected a wide distribution of demographics, with a mean age of 63.2 (SD 18.0) years at the time of the ECG. In the development population, 1 393 683 (48.2%) of the ECGs were from women, 1 904 662 (65.9%) ECGs were from non-Hispanic White, 437 244 (15.1%) non-Hispanic Black, 298 490 (10.3%) Hispanic, 34 855 (1.2%) non-Hispanic Asian, and 213 133 (7.4%) from patients from other racial backgrounds (see Supplementary material online, *Table S1*), reflecting broad and diverse representation. In a period when linkage to hospital admission records data was possible (with the digitization of health records, starting in 2013), 37.5%, 13.5%, and

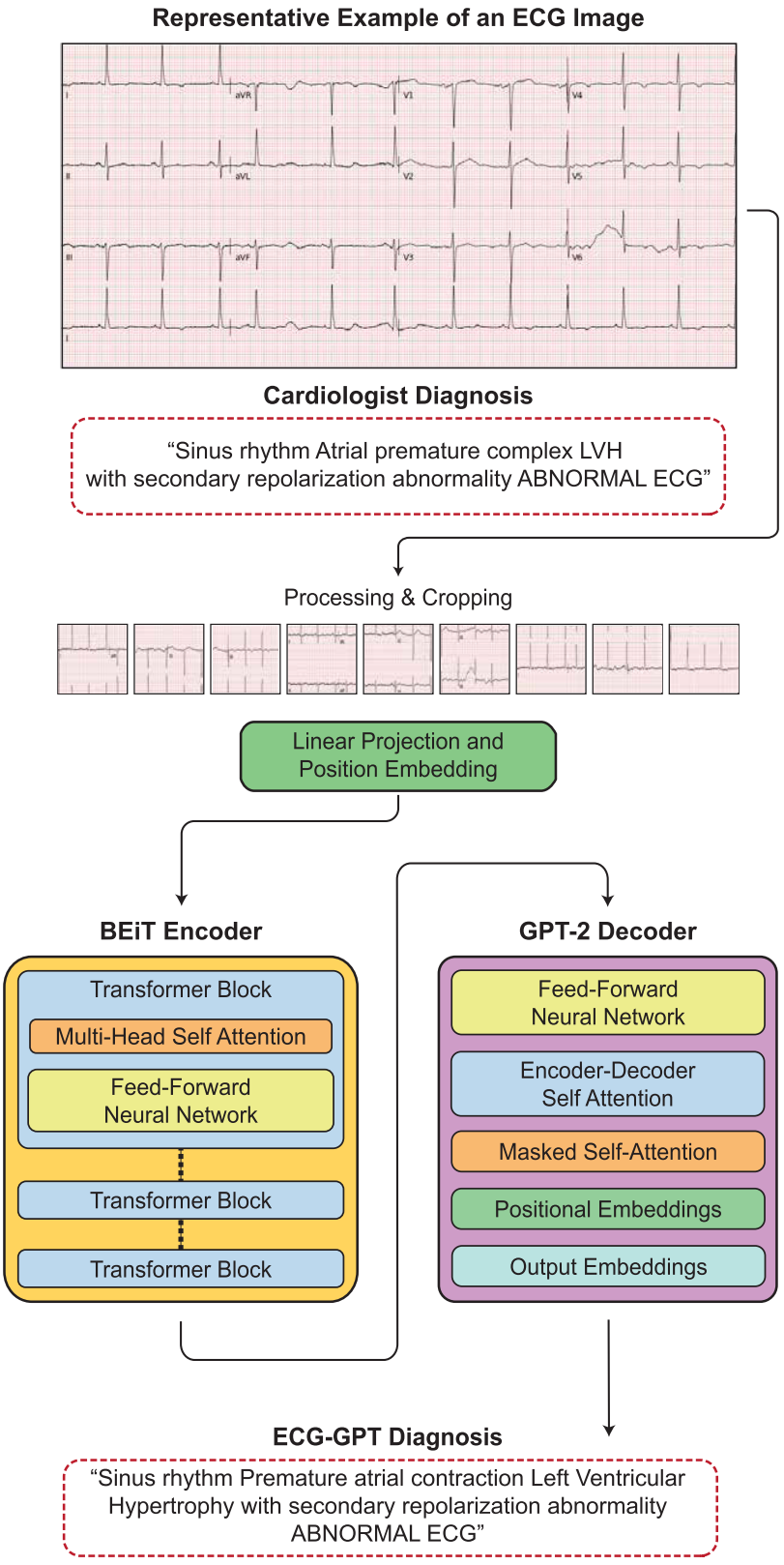


Figure 1 Vision encoder-decoder model architecture and sample ECG-GPT output interpretation statement. Abbreviations: ECG, electrocardiogram; BEiT, Bidirectional Encoder representation from Image Transformers; GPT-2, Generative Pretrained Transformer-2.

49.0% of ECGs were performed in the inpatient, emergency, and outpatient settings, respectively.

The standardized diagnosis statements corresponding to each ECG had a median length of 81 characters (IQR 51, 120) and a median length of 12 tokens (IQR 8, 17) after tokenization. In the development cohort, most ECGs reported sinus rhythm (1,935,574, 67.0%), followed by left atrial enlargement (359,747, 12.5%), left ventricular hypertrophy (342,340, 11.9%), and AF (281,961, 9.8%). A total of 265 640 (9.2%) and 126 757 (4.4%) ECGs reported ST and SB, respectively. RBBB and LBBB were present in 195 474 (6.8%) and 94 928 (3.3%) ECGs, respectively, and 211 836 (7.33%) ECGs reported first-degree atrioventricular block (1dAVb). 20 003 (0.69%) of ECGs report acute myocardial infarction (AMI). Though the presence of multiple rhythms over a 10-second duration would be unusual, if the patient had a series of conduction and rhythm disorders present together (such as AF with premature ventricular contractions and bundle branch blocks), all would be included in the statement and thus the extracted labels. 104 451 (3.6%) ECGs featured multiple key rhythm disorders, including AF, ST, SB, premature atrial contractions (PACs), and premature ventricular contractions (PVCs). Complete demographic information and the prevalences for each of the 26 extracted conditions in the development set, the held-out test set, and the external validation sets from MSHS and HMH are listed in Supplementary material online, *Table S1*. Demographic information and prevalences for a limited set of conditions for the remaining external validation sets is listed in Supplementary material online, *Table S2*.

Semantic similarity

Using a fine-tuned DistilBERT language model, we generated embeddings for each reference and model-generated statement for the 250 589 ECGs in the internal held-out test set, obtained from 57 860 patients who had not contributed any data to the development set (see Supplementary material online, *Table S1*). The median cosine similarity between the embeddings for reference statements and their paired model-generated statements was 0.90 (IQR 0.83–0.97). This result was significantly higher than the median cosine similarity of 0.73 (IQR 0.67–0.78, $P < 0.001$) between 100 000 randomly selected combinations of reference and model-generated statements.

In a secondary analysis to assess the model's ability to capture individual conditions within their full clinical context, the median pairwise cosine similarity was significantly greater than the respective median random cosine similarity across all 26 subsets (*Table 1*). Across conditions, pairwise and baseline similarities ranged from 0.84 to 0.97 and 0.72 to 0.84, respectively. For key rhythm disorders—AF, ST, SB, PACs, and PVCs—pairwise and random similarities ranged from 0.91 to 0.93 and 0.77 to 0.81, respectively. For conduction abnormalities—LBBB, RBBB, 1dAVb, left anterior fascicular block (LAFB), and left posterior fascicular block (LPFB)—pairwise and random similarities ranged from 0.91 to 0.95 and 0.79 to 0.84, respectively.

Syntactic similarity

For conventional natural language generation metrics comparing reference and model-generated interpretations for the held-out test set, we report scores of 0.663 and 0.655 for ROUGE-1 and ROUGE-L, respectively. For BLEU scores, we report scores ranging from 0.595 for BLEU-1 to 0.419 for BLEU-4. We also report METEOR and CIDEr scores of 0.693 and 3.32, respectively (see Supplementary material online, *Table S3*).

Structured label assessment

Model performance in the held-out test set for each of the 26 extracted labels is recorded in *Table 2*. A performance comparison with two previously published multilabel CNN models is recorded in Supplementary material online, *Table S4*. For AF, ST, SB, PACs, and PVCs, AUROCs and AUPRCs ranged from 0.80–0.95 and 0.50–0.86, respectively. For LBBB, RBBB, 1dAVb, LAFB, and LPFB, the AUROCs and AUPRCs ranged from 0.88–0.96 and 0.23–0.86, respectively. For AMI, the model had an AUROC and AUPRC of 0.85 and 0.18, respectively. Across all 26 conditions, diagnostic accuracy ranged between 0.93 and 0.99. Moreover, the model performed comparably across patient subgroups (see Supplementary material online, *Table S5*). There was no significant difference in the performance for the labels of interest when comparing ECGs recorded in the inpatient, emergency, and outpatient settings, with median AUROCs of 0.89 (IQR 0.85–0.92), 0.90 (IQR 0.87–0.94), and 0.89 (IQR 0.83–0.94), respectively ($P = 0.696$).

External validation—Mount Sinai Health System

We used 1 434 455 ECG images generated from raw 12-lead ECG signal data and 10 052 prospectively acquired native ECG images from MSHS for validation. In assessing semantic similarity for the generated ECG images, embeddings had a median pairwise similarity of 0.86 (IQR 0.79–0.96), significantly greater than the median baseline similarity of 0.74 among random pairings (IQR 0.69–0.80, $P < 0.001$). For the native ECG images, embeddings had a median pairwise and baseline similarity of 0.80 (IQR 0.72–0.86) and 0.73 (IQR 0.68–0.78), respectively. This separation persisted across the 26 subsets corresponding to each extracted label (see Supplementary material online, *Table S6* and *Table S7*).

The model performed well in clinical assessment across the 26 extracted labels. For AF, ST, SB, PACs, and PVCs, AUROCs and AUPRCs ranged from 0.74 to 0.94 and 0.40 to 0.83, respectively, for the generated ECG images and 0.67–0.91 and 0.30–0.76, respectively, for the native ECG images. For LBBB, RBBB, 1dAVb, LAFB, and LFPB, AUROCs and AUPRCs ranged from 0.87–0.97 and 0.14–0.83, respectively, for the generated ECG images and 0.71–0.97 and 0.05–0.63, respectively, for the native ECG images. For AMI, AUROCs, and AUPRCs, respectively, were 0.83 and 0.15 for generated ECG images and 0.82 and 0.03 for native ECG images. Model performance for a set of key conditions in the MSHS validation datasets and the HMH validation dataset is reported in *Table 3*. Performance across all extracted labels for the MSHS-generated and native ECG images is reported in Supplementary material online, *Table S8* and *Table S9*, respectively.

External validation—Houston Methodist Hospital

We also deployed ECG-GPT in a set of 62 876 native ECG images from HMH. Across conditions, embeddings had a median pairwise similarity of 0.81 (IQR 0.75–0.88), significantly greater than the median baseline similarity of 0.77 among 2 random statements (IQR 0.73–0.82, $P < 0.001$). This separation persisted across the majority of the 26 subsets (see Supplementary material online, *Table S10*). In clinical assessment, AUROCs and AUPRCs, respectively, for AF, ST, SB, PACs, and PVCs ranged from 0.65–0.96 and 0.13–0.68. For LBBB, RBBB, 1dAVb, LAFB, and LFPB, AUROCs and AUPRCs ranged from 0.74–0.97 and 0.04–0.77, respectively. AUROCs and AUPRCs for AMI were 0.89 and 0.02, respectively. Model performance across all 26 conditions is reported in Supplementary material online, *Table S11*.

Table 1 Pairwise and baseline similarity between reference and model-generated ECG interpretation statements in the held-out test set

Labels	Pairwise similarity	Baseline similarity	*P*-Value
Overall	0.896 (0.827–0.970)	0.727 (0.671–0.783)	<0.001
Sinus Rhythm	0.884 (0.815–0.968)	0.746 (0.691–0.800)	<0.001
AF	0.914 (0.859–0.968)	0.794 (0.748–0.839)	<0.001
Atrial Flutter	0.909 (0.844–0.961)	0.782 (0.740–0.839)	<0.001
ST	0.913 (0.849–0.975)	0.788 (0.739–0.837)	<0.001
SB	0.945 (0.871–1.000)	0.807 (0.755–0.860)	<0.001
Sinus Arrythmia	0.969 (0.911–1.000)	0.819 (0.743–0.884)	<0.001
LBBB	0.929 (0.846–1.000)	0.824 (0.768–0.880)	<0.001
RBBB	0.939 (0.885–0.994)	0.826 (0.778–0.874)	<0.001
LAFB	0.931 (0.875–0.975)	0.803 (0.756–0.850)	<0.001
LPFB	0.949 (0.897–0.989)	0.842 (0.787–0.891)	<0.001
SVT	0.899 (0.828–0.955)	0.778 (0.726–0.828)	<0.001
PAC	0.908 (0.848–0.956)	0.769 (0.715–0.821)	<0.001
PVC	0.927 (0.874–0.969)	0.784 (0.736–0.830)	<0.001
LAE	0.914 (0.841–0.968)	0.766 (0.713–0.822)	<0.001
RAE	0.920 (0.866–0.964)	0.770 (0.718–0.823)	<0.001
LVH	0.935 (0.879–0.979)	0.799 (0.745–0.853)	<0.001
RVH	0.954 (0.882–0.987)	0.794 (0.736–0.858)	<0.001
Low Voltage	0.914 (0.845–0.968)	0.756 (0.698–0.816)	<0.001
Left Axis Deviation	0.913 (0.851–0.969)	0.760 (0.712–0.810)	<0.001
Acute MI	0.902 (0.840–0.958)	0.764 (0.717–0.819)	<0.001
Lead Reversal	0.923 (0.844–0.969)	0.816 (0.732–0.889)	<0.001
1st degree AVb	0.913 (0.849–0.967)	0.787 (0.736–0.840)	<0.001
2nd degree AVb	0.898 (0.837–0.948)	0.815 (0.759–0.865)	<0.001
3rd degree AVb	0.870 (0.801–0.929)	0.753 (0.690–0.807)	<0.001
WPW	0.837 (0.719–0.935)	0.724 (0.670–0.838)	<0.001
Repol Abnormality	0.922 (0.858–0.971)	0.780 (0.726–0.840)	<0.001

Abbreviations: AF, atrial fibrillation; ST, sinus tachycardia; SB, sinus bradycardia; LBBB, left bundle branch block; RBBB, right bundle branch block; LAFB, left anterior fascicular block; LPFB, left posterior fascicular block; SVT, supraventricular tachycardia; PAC, premature atrial complexes; PVC, premature ventricular complexes; LAE, left atrial enlargement; RAE, right atrial enlargement; LVH, left ventricular hypertrophy; RVH, right ventricular hypertrophy; MI, myocardial infarction; AVb, atrioventricular block; WPW, Wolff-Parkinson-White syndrome; Repol, repolarization. *Acute MI includes ST Elevation MI (STEMI).

External validation—Lake Regional Hospital

Next, we used a real-world dataset of 64 ECG images collected at LRH in Osage Beach, MO. In these ECGs, the model had AUROCs of 0.95, 0.95, and 0.75 for AF, ST, and SB, respectively. For LBBB, RBBB, and AVb, the model reported AUROCs of 0.82, 1.00, and 0.87, respectively.

External validation—open-source datasets

The model's diagnostic performance across the four open-source datasets used for validation is noted in *Table 4*. First, in 2 228 236 ECGs from the CODE15 dataset collected by the Telehealth Network of Minas Gerais (TNMG), Brazil, between 2010 and 2017.[12,23] Here, AUROCs for rhythm disorders were 0.94, 0.92, and 0.92, and AUPRCs were 0.56, 0.54, and 0.11, for AF, ST, and SB, respectively. The model performed similarly for conduction abnormalities, with AUROCs of 0.90, 0.96, and 0.89, and AUPRCs of 0.62, 0.67, and 0.29 for LBBB, RBBB, and AVb, respectively. When deployed to a smaller, cardiologist-validated dataset collected by TNMG in Brazil between April and September 2018,[15] AUROCs were higher across nearly all diagnostic labels. The model reported AUROCs of 1.00 (rounded from 0.998), 0.98, and 0.94 for AF, ST, and SB, respectively. For LBBB, RBBB, and AVb, the model reported AUROCs of 0.94, 0.97, and 0.85, respectively.

The third of these four open-source datasets consisted of 45 389 ECGs obtained from patients enrolled in the UK Biobank. Here, the model had AUROCs ranging from 0.91 to 0.99, thus highlighting the reproducibility of our approach across clinical and nonclinical settings. Lastly, in 21 784 ECGs from the Germany-based PTB-XL dataset, the model had AUROCs ranging from 0.84–0.98.

Discussion

We describe the development and external validation of ECG-GPT, a first-of-its-kind AI pipeline that enables the direct generation of automated, complete ECG interpretations from images of ECGs in any format. The model performs well against clinician-certified reports across various natural language generation metrics and diagnostic labels spanning a wide array of conduction, rhythm, and structural heart disorders. The model's scalability is further supported by its robust performance across a range of demographically, temporally, and geographically distinct cohorts, its online

Table 2 Clinical assessment of model-generated ECG interpretation statements in the held-out test set

Labels	PPV	NPV	Specificity	Sensitivity	Accuracy	AUROC	AUPRC	F1
Sinus Rhythm	0.971	0.913	0.944	0.954	0.950	0.949 (0.948–0.95)	0.956 (0.955–0.957)	0.962
AF	0.853	0.984	0.983	0.862	0.970	0.922 (0.92–0.925)	0.75 (0.745–0.755)	0.858
Atrial Flutter	0.381	0.997	0.970	0.868	0.968	0.919 (0.914–0.923)	0.333 (0.324–0.342)	0.529
ST	0.936	0.992	0.994	0.912	0.988	0.953 (0.951–0.955)	0.86 (0.856–0.865)	0.924
SB	0.902	0.992	0.995	0.867	0.988	0.931 (0.928–0.934)	0.79 (0.783–0.797)	0.885
Sinus Arrythmia	0.395	0.998	0.964	0.922	0.963	0.943 (0.94–0.947)	0.366 (0.358–0.373)	0.553
LBBB	0.853	0.995	0.995	0.855	0.990	0.925 (0.921–0.929)	0.735 (0.725–0.744)	0.854
RBBB	0.909	0.995	0.993	0.935	0.989	0.964 (0.962–0.965)	0.854 (0.849–0.859)	0.922
LAFB	0.701	0.989	0.980	0.808	0.971	0.894 (0.891–0.898)	0.577 (0.568–0.583)	0.751
LPFB	0.271	0.999	0.988	0.861	0.987	0.924 (0.915–0.933)	0.234 (0.222–0.248)	0.412
SVT	0.399	0.999	0.996	0.698	0.994	0.847 (0.833–0.861)	0.28 (0.259–0.303)	0.508
PAC	0.771	0.976	0.989	0.616	0.967	0.802 (0.798–0.806)	0.498 (0.49–0.506)	0.685
PVC	0.799	0.988	0.983	0.848	0.973	0.915 (0.913–0.918)	0.688 (0.681–0.694)	0.823
LAE	0.679	0.961	0.956	0.706	0.927	0.831 (0.828–0.834)	0.513 (0.508–0.519)	0.692
RAE	0.251	0.999	0.990	0.764	0.989	0.877 (0.864–0.889)	0.192 (0.178–0.207)	0.377
LVH	0.729	0.983	0.960	0.864	0.949	0.912 (0.910–0.914)	0.645 (0.641–0.651)	0.791
RVH	0.118	0.998	0.928	0.822	0.927	0.875 (0.868–0.882)	0.099 (0.095–0.104)	0.206
Low Voltage	0.660	0.977	0.963	0.762	0.945	0.862 (0.86–0.865)	0.524 (0.518–0.531)	0.708
Left Axis Deviation	0.649	0.975	0.959	0.753	0.940	0.856 (0.853–0.859)	0.511 (0.504–0.518)	0.697
Acute MI[a]	0.255	0.998	0.986	0.711	0.985	0.849 (0.838–0.86)	0.183 (0.171–0.194)	0.375
Lead Reversal	0.325	1.000	0.997	0.718	0.997	0.858 (0.836–0.879)	0.234 (0.206–0.261)	0.447
1st degree AVb	0.743	0.983	0.978	0.787	0.964	0.883 (0.88–0.886)	0.601 (0.594–0.608)	0.764
2nd degree AVb	0.379	0.999	0.999	0.643	0.998	0.821 (0.796–0.846)	0.244 (0.207–0.283)	0.477
3rd degree AVb	0.260	1.000	0.998	0.693	0.998	0.846 (0.816–0.875)	0.181 (0.149–0.21)	0.378
WPW	0.444	1.000	1.000	0.624	0.999	0.812 (0.771–0.853)	0.277 (0.218–0.353)	0.519
Repol Abnormality	0.515	0.973	0.951	0.661	0.930	0.806 (0.802–0.809)	0.365 (0.359–0.37)	0.579

Abbreviations: PPV, Positive Predictive Value; NPV, Negative Predictive Value; AUROC, area under the receiver operator characteristic; AUPRC, area under precision-recall curve; AF, atrial fibrillation; ST, sinus tachycardia; SB, sinus bradycardia; LBBB, left bundle branch block; RBBB, right bundle branch block; LAFB, left anterior fascicular block; LPFB, left posterior fascicular block; SVT, supraventricular tachycardia; PAC, premature atrial complexes; PVC, premature ventricular complexes; LAE, left atrial enlargement; RAE, right atrial enlargement; LVH, left ventricular hypertrophy; RVH, right ventricular hypertrophy; MI, myocardial infarction; AVb, atrioventricular block; WPW, Wolff-Parkinson-White syndrome; Repol, repolarization.
[a]Acute MI includes ST Elevation MI (STEMI).

deployment, and the capacity to containerize and deploy the model without sharing data in a federated approach.

Our work on format-independent, image-based ECG captioning represents a novel development for vision-text machine learning models. For conventional natural language generation metrics, ECG-GPT matches or outperforms most state-of-the-art medical image captioning models.[24,25] We report CIDEr and METEOR scores of 4.69 and 0.75, respectively, compared with scores of 2.55 and 0.27 reported in a previous study for generating free-text reports from ECG signals, suggesting high consistency between reference and generated reports.[26] Machine learning-based multilabel models have been previously developed to simultaneously diagnose large sets of conditions, but these approaches are inherently limited to those labels selected for training and do not capture the full diagnostic range of ECGs.[12,15,27] ECG-GPT performed comparably to two previously published multilabel CNN models[14,27] across common labels. Moreover, the utility of such signal models is limited to healthcare systems with the resources to store signal data and incorporate models into the clinical workflow. We demonstrate consistent performance across a range of external validation sets for classifying key rhythm and conduction disorders. Of note, the performance for labels in external validation sets where specific labels were explicitly available matches the performance on those labels in prior published reports. The simplicity of this system, based on images, is that there is inherent interoperability and no requirement to integrate with ECG machines to extract signals. This approach is particularly advantageous in low-resource regions, where ECGs are currently not stored beyond printing ECG images at the time of acquisition.[28] This approach also adds convenience and provides access in any venue, including, for example, to emergency medical services providers or in remote locations.

ECG-GPT can be used directly by clinicians at the point of care by uploading ECG images from their phones or as scanned images to a web-based interface, with a demonstration accompanying this study.[29] This applies anywhere that end-users may still lack access to automated reads or require interpretation before a specialist's review. The image-based model can also be more easily integrated into repositories of scanned ECGs, the most prevalent and interoperable format for storing and sharing ECGs. Further, ECG-GPT has a unique combination of diagnostic accuracy and range, demonstrating expert-level performance for key conditions while also retaining the capability to generate statements for rare conditions frequently

Table 3 Clinical assessment of model-generated ECG interpretation statements in external validation sets with corresponding diagnosis statements

	Labels	PPV	NPV	Specificity	Sensitivity	Accuracy	AUROC	AUPRC	F1
MSHS Generated Image Dataset	AF	0.693	0.986	0.971	0.821	0.960	0.896 (0.882–0.910)	0.582 (0.550–0.615)	0.752
	ST	0.910	0.988	0.989	0.898	0.979	0.944 (0.943–0.944)	0.828 (0.827–0.830)	0.904
	SB	0.948	0.967	0.995	0.734	0.965	0.865 (0.863–0.866)	0.726 (0.724–0.728)	0.828
	PAC	0.763	0.968	0.991	0.479	0.961	0.735 (0.733–0.737)	0.396 (0.392–0.399)	0.589
	PVC	0.749	0.988	0.981	0.826	0.972	0.904 (0.902–0.905)	0.629 (0.626–0.633)	0.785
	LBBB	0.746	0.998	0.992	0.914	0.990	0.953 (0.952–0.954)	0.684 (0.679–0.688)	0.821
	RBBB	0.876	0.996	0.990	0.940	0.987	0.965 (0.965–0.966)	0.827 (0.825–0.829)	0.907
	1dAVb	0.699	0.990	0.977	0.845	0.970	0.911 (0.910–0.912)	0.599 (0.596–0.603)	0.765
	LAFB	0.558	0.991	0.977	0.776	0.969	0.877 (0.875–0.878)	0.441 (0.437–0.445)	0.649
	LPFB	0.179	0.999	0.986	0.753	0.985	0.869 (0.864–0.875)	0.136 (0.131–0.141)	0.289
MSHS Native Image Dataset	AF	0.693	0.986	0.971	0.821	0.96	0.896 (0.882–0.910)	0.582 (0.550–0.615)	0.752
	ST	0.797	0.978	0.977	0.804	0.959	0.890 (0.878–0.903)	0.661 (0.634–0.685)	0.800
	SB	0.895	0.978	0.988	0.822	0.970	0.905 (0.894–0.916)	0.755 (0.732–0.781)	0.857
	PAC	0.738	0.957	0.992	0.347	0.950	0.669 (0.651–0.688)	0.298 (0.264–0.336)	0.472
	PVC	0.694	0.986	0.975	0.801	0.963	0.888 (0.873–0.903)	0.570 (0.539–0.605)	0.744
	LBBB	0.559	0.997	0.981	0.898	0.979	0.939 (0.921–0.958)	0.504 (0.450–0.555)	0.689
	RBBB	0.645	0.998	0.960	0.973	0.961	0.966 (0.960–0.973)	0.629 (0.597–0.655)	0.776
	1dAVb	0.550	0.988	0.949	0.843	0.942	0.896 (0.882–0.910)	0.475 (0.441–0.508)	0.666
	LAFB	0.415	0.979	0.976	0.453	0.956	0.714 (0.689–0.740)	0.208 (0.176–0.246)	0.433
	LPFB	0.084	0.999	0.978	0.625	0.977	0.802 (0.716–0.887)	0.054 (0.023–0.086)	0.148
HMH Native Image Dataset	AF	0.801	0.995	0.994	0.837	0.989	0.916 (0.907–0.924)	0.675 (0.652–0.698)	0.819
	ST	0.596	0.998	0.978	0.937	0.976	0.957 (0.952–0.962)	0.561 (0.542–0.579)	0.729
	SB	0.977	0.917	0.997	0.589	0.923	0.793 (0.788–0.797)	0.650 (0.642–0.656)	0.735
	PAC	0.798	0.978	0.997	0.308	0.976	0.653 (0.642–0.663)	0.267 (0.249–0.289)	0.444
	PVC	0.139	0.995	0.787	0.892	0.791	0.840 (0.833–0.846)	0.128 (0.123–0.133)	0.241
	LBBB	0.767	0.999	0.997	0.905	0.996	0.951 (0.940–0.962)	0.696 (0.670–0.730)	0.831
	RBBB	0.809	0.998	0.991	0.944	0.989	0.967 (0.963–0.972)	0.765 (0.752–0.780)	0.871
	1dAVb	0.809	0.988	0.991	0.755	0.980	0.873 (0.866–0.881)	0.622 (0.606–0.642)	0.781
	LAFB	0.451	0.993	0.981	0.698	0.975	0.840 (0.827–0.852)	0.321 (0.298–0.345)	0.548
	LPFB	0.082	0.998	0.984	0.497	0.982	0.740 (0.704–0.777)	0.042 (0.030–0.056)	0.141

Abbreviations: MSHS, Mount Sinai Health System; HMH, Houston Methodist Hospital; PPV, Positive Predictive Value; NPV, Negative Predictive Value; AUROC, area under the receiver operator characteristic; AUPRC, area under precision-recall curve; AF, atrial fibrillation; ST, sinus tachycardia; SB, sinus bradycardia; PAC, premature atrial complexes; PVC, premature ventricular complexes; LBBB, left bundle branch block; RBBB, right bundle branch block; 1dAVb, first-degree atrioventricular block; LAFB, left anterior fascicular block; LPFB, left posterior fascicular block.

not captured by standard multi-label models. This feature could make ECG-GPT a tool for generating pre-reads and enabling more efficient triage globally in areas with insufficient access to specialists and computerized ECG interpretation. The study here focuses on the robustness of ECG-GPT in generating accurate interpretations from ECG images, but future prospective studies are warranted to fully assess its integration into clinical workflows in care settings where the tool is used.

Our study has several limitations. First, while the model accurately diagnosed the selected conditions, it is impossible to determine and thus evaluate the performance for the full extent of possible diagnoses the model could output for a given ECG image due to the size and variety of the corpus of diagnosis statements used for model development. However, we report model performance across various rhythm and conduction disorders of varying severity and prevalence, suggesting that ECG-GPT's performance generalizes to other conditions. Additionally, though model performance was evaluated for the full range of 26 selected conditions in the datasets from YNHH, MSHS, and HMH, only a select set of 6 labels was available for LRH and for each international validation site, limiting the evaluation of model performance in global settings for a broader set of diagnostic labels. Moreover, using the federated approach implemented for external validation, ECG-GPT could be continually fine-tuned to improve performance in individual healthcare systems. This would ensure a reliable pipeline with consistent performance for future ECGs within the specific patient populations in which the model is deployed. Second, it currently only generates interpretation statements in English, which may limit its utility in non-English-speaking settings.

Third, while four different formats were used during model development, we cannot ascertain whether the model generalizes equally well to every other novel ECG image format. However, the model's performance within the large, native

Table 4 Clinical assessment of model-generated ECG interpretation statements on external validation sets with limited diagnostic labels

	Labels	PPV	NPV	Specificity	Sensitivity	Accuracy	AUROC	AUPRC	F1
Cardiologist Validated CODE15	AF	0.765	1.000	0.995	1.000	0.996	0.998 (0.996–1.00)	0.765 (0.550–0.947)	0.867
	ST	0.884	0.999	0.994	0.974	0.993	0.984 (0.959–1)	0.862 (0.749–0.973)	0.927
	SB	0.139	1.000	0.888	1.000	0.890	0.944 (0.934–0.954)	0.139 (0.077–0.202)	0.244
	LBBB	1.000	0.995	1.000	0.871	0.996	0.935 (0.876–0.995)	0.875 (0.759–0.971)	0.931
	RBBB	0.829	0.998	0.992	0.944	0.990	0.968 (0.93–1)	0.785 (0.639–0.898)	0.883
	AVb	0.629	0.990	0.985	0.710	0.976	0.847 (0.766–0.929)	0.456 (0.293–0.603)	0.667
CODE15	AF	0.618	0.998	0.990	0.897	0.988	0.943 (0.942–0.945)	0.556 (0.551–0.559)	0.731
	ST	0.625	0.997	0.989	0.855	0.986	0.922 (0.92–0.923)	0.538 (0.534–0.543)	0.722
	SB	0.109	0.999	0.872	0.972	0.874	0.922 (0.921–0.923)	0.107 (0.106–0.108)	0.197
	LBBB	0.766	0.997	0.996	0.810	0.993	0.903 (0.901–0.905)	0.624 (0.619–0.629)	0.788
	RBBB	0.724	0.998	0.990	0.923	0.988	0.956 (0.955–0.957)	0.670 (0.667–0.673)	0.811
	AVb	0.353	0.997	0.977	0.811	0.974	0.894 (0.892–0.896)	0.289 (0.286–0.293)	0.492
UK Biobank	AF	0.939	0.998	0.999	0.876	0.997	0.937 (0.925–0.95)	0.824 (0.799–0.855)	0.906
	ST	0.862	1.000	1.000	0.815	0.999	0.907 (0.868–0.947)	0.703 (0.599–0.798)	0.838
	SB	0.973	0.894	0.981	0.853	0.924	0.917 (0.914–0.919)	0.894 (0.890–0.898)	0.909
	LBBB	0.871	1.000	0.999	0.982	0.999	0.991 (0.984–0.997)	0.855 (0.822–0.884)	0.923
	RBBB	0.832	1.000	0.996	0.979	0.996	0.987 (0.983–0.992)	0.814 (0.791–0.840)	0.899
	AVb	0.663	0.995	0.973	0.911	0.969	0.942 (0.936–0.947)	0.609 (0.592–0.627)	0.767
PTB-XL	AF	0.787	0.995	0.981	0.931	0.978	0.956 (0.949–0.962)	0.738 (0.716–0.758)	0.853
	ST	0.836	0.996	0.993	0.907	0.99	0.950 (0.940–0.960)	0.762 (0.732–0.787)	0.870
	SB	0.184	0.998	0.874	0.942	0.876	0.908 (0.899–0.917)	0.175 (0.161–0.187)	0.308
	LBBB	0.857	0.999	0.996	0.944	0.995	0.970 (0.960–0.980)	0.811 (0.783–0.846)	0.899
	RBBB	0.691	0.999	0.989	0.980	0.989	0.984 (0.978–0.99)	0.678 (0.645–0.713)	0.810
	AVb	0.419	0.989	0.963	0.711	0.954	0.837 (0.821–0.853)	0.309 (0.284–0.335)	0.528
LRH	AFIB	0.923	0.980	0.980	0.923	0.969	0.952 (0.874–1.00)	0.868 (0.665–1.00)	0.923
	ST	1.000	0.982	1.000	0.900	0.984	0.950 (0.852–1.00)	0.916 (0.714–1.00)	0.947
	SB	1.000	0.915	1.000	0.500	0.922	0.750 (0.587–0.913)	0.578 (0.245–0.873)	0.667
	LBBB	1.000	0.930	1.000	0.636	0.938	0.818 (0.669–0.967)	0.699 (0.427–0.873)	0.778
	RBBB	1.000	1.000	1.000	1.000	1.000	1.00 (1.00–1.00)	1.00 (1.00–1.00)	1.000
	AVb	0.692	0.961	0.925	0.818	0.906	0.871 (0.747–0.996)	0.598 (0.312–0.818)	0.750

Abbreviations: ECG, electrocardiogram; PPV, Positive Predictive Value; NPV, Negative Predictive Value; Spec, specificity; Sens, sensitivity; AUROC, area under the receiver operator characteristic; AUPRC, area under precision-recall curve. AF, atrial fibrillation; ST, sinus tachycardia; SB, sinus bradycardia; LBBB, left bundle branch block; RBBB, right bundle branch block; AVb, atrioventricular block.

image datasets from MSHS, HMH, and LRH, each consisting of ECG images plotted in configurations distinct from those used in model development, indicates the model can generate accurate interpretation statements for ECG images with variations not seen during training. Finally, though strictly expert-validated diagnosis statements were used to develop the model, these statements are not always completely accurate, limiting the model's performance. As evidenced by the high diagnostic accuracy of the model in the external set of ECGs manually annotated by two cardiologists, the model may perform better if developed and evaluated in more rigorously validated diagnosis statements. Furthermore, the current practice of clinicians overreading and correcting computer-generated reads without version control precludes the head-to-head assessment of ECG-GPT against computer-generated reads. Future studies could compare the automated outputs of the model to ECGs labeled by cardiologists without prior computerized interpretation, providing a clearer assessment of the model's performance relative to human experts. Nevertheless, the higher performance in labels assigned by more than one expert suggests that it likely performs at or above the performance of the current computerized reads at the US health systems, especially for the tested diagnoses.

We have developed and extensively validated a novel vision-text transformer capable of generating complete diagnostic statements from ECG images in any lead layout and configuration. Our approach represents a scalable and accessible strategy for generating accurate, expert-level reports from photos of ECGs, enabling accurate ECG interpretation anywhere that an ECG image, paper, or digital can be produced.

Supplementary material

Supplementary material is available at *European Heart Journal – Digital Health*.

Acknowledgements

RK conceived the study and accessed the data. AK, VS, and RK developed the model. AK, VS, AV, and RK pursued the statistical

analysis. AK, VS, and LSD drafted the manuscript. All authors provided feedback regarding the study design and made critical contributions to the manuscript writing. RK supervised the study, procured funding, and is the guarantor.

Author contributions

Akshay Khunte (BS (Conceptualization [supporting]; Data curation [supporting]; Formal analysis [lead]; Investigation [lead]; Methodology [lead]; Validation [lead]; Visualization [lead]; Writing—original draft [lead]; Writing—review & editing [lead])), Veer Sangha (BS (Formal analysis [supporting]; Methodology [supporting]; Validation [supporting]; Visualization [supporting]; Writing—original draft [supporting]; Writing—review & editing [supporting])), Evangelos K Oikonomou (MD DPhil (Data curation [supporting]; Formal analysis [supporting]; Investigation [supporting]; Methodology [supporting]; Writing—original draft [supporting]; Writing—review & editing [supporting])), Lovedeep S Dhingra (MBBS MHS (Formal analysis [supporting]; Writing—original draft [supporting]; Writing—review & editing [supporting])), Arya Aminorroaya (MD MPH (Writing—original draft [supporting]; Writing—review & editing [supporting])), Andreas Coppi (PhD (Data curation [supporting]; Formal analysis [supporting]; Methodology [supporting]; Validation [supporting])), Sumukh Vasisht Shankar (MS (Formal analysis [supporting]; Validation [supporting])), Elijah Rockers {MS [Validation (supporting)]}, Bobak J Mortazavi {PhD [Methodology (supporting)]}, Deepak L Bhatt {MD MPH [Validation (supporting)]}, Harlan M Krumholz (MD SM (Conceptualization [supporting]; Writing—review & editing [supporting])), Sadeer Al-Kindi {MD [Validation (supporting)]}, Girish N Nadkarni (MD MPH (Formal analysis [supporting]; Validation [supporting]; Writing—review & editing [supporting])), Akhil Vaid (MD (Formal analysis [supporting]; Validation [supporting]; Writing—review & editing [supporting])), and Rohan Khera (MD MS (Conceptualization [lead]; Data curation [lead]; Formal analysis [supporting]; Funding acquisition [lead]; Investigation [lead]; Methodology [supporting]; Project administration [lead]; Resources [lead]; Supervision [lead]; Validation [supporting]; Visualization [supporting]; Writing—original draft [supporting]; Writing—review & editing [supporting]))

Funding

This study was supported by research funding awarded to Dr. Khera by the Yale School of Medicine and grant support from the National Institutes of Health (under awards R01HL167858, R01AG089981, and K23HL153775) and the Doris Duke Charitable Foundation (under award, 2022060). Dr. Oikonomou received support from the National Heart, Lung, and Blood Institute of the National Institutes of Health (under award 1F32HL170592). The funders had no role in the design and conduct of the study; collection, management, analysis, and interpretation of the data; preparation, review, or approval of the manuscript; and decision to submit the manuscript for publication.

Conflict of interest: Mr. Khunte, Mr. Sangha, and Dr. Khera are the coinventors of U.S. Provisional Patent Application No. 63/428 569. Mr. Sangha and Dr. Khera are the coinventors of U.S. Pending Patent Application No. 63/346,610, and are co-founders of Ensight-AI with Dr. Krumholz. Dr. Khera is the coinventor of U.S. Provisional Patent Application No. 63/177 117 (unrelated to current work) and is a co-founder of Evidence2Health, a precision health platform for evidence-based care. He is also an associate editor at JAMA and received support from the National Institutes of Health (under award R01HL167858, R01AG089981, and K23HL153775) and the Doris Duke Charitable Foundation (under award, 2022060). He also receives research support, through Yale, from Bristol-Myers Squibb, Novo Nordisk, and BridgeBio. Dr. Oikonomou receives support from the National Heart, Lung, and Blood Institute of the National Institutes of Health (under award F32HL170592). He is a co-founder of Evidence2Health LLC, a co-inventor in patent applications (18/813,882, 17/720,068, 63/508,315, 63/580,137, 63/619,241, 63/562,335, US12067714B2, US11948230B2), has been a consultant for Caristo Diagnostics Ltd and Ensight-AI Inc, and has received royalty fees from technology licensed through the University of Oxford. He also serves as Associate Editor for European Heart Journal. Dr. Nadkarni is a founder of Renalytix, Pensieve, and Verici and provides consultancy services to AstraZeneca, Reata, Renalytix, Siemens Healthineer, and Variant Bio, and serves a scientific advisory board member for Renalytix and Pensieve. He also has equity in Renalytix, Pensieve, and Verici. Dr. Krumholz works under contract with the Centers for Medicare & Medicaid Services to support quality measurement programs, was a recipient of a research grant from Johnson & Johnson, through Yale University, to support clinical trial data sharing; was a recipient of a research agreement, through Yale University, from the Shenzhen Center for Health Information for work to advance intelligent disease prevention and health promotion; collaborates with the National Center for Cardiovascular Diseases in Beijing; receives payment from the Arnold & Porter Law Firm for work related to the Sanofi clopidogrel litigation, from the Martin Baughman Law Firm for work related to the Cook Celect IVC filter litigation, and from the Siegfried and Jensen Law Firm for work related to Vioxx litigation; chairs a Cardiac Scientific Advisory Board for UnitedHealth; was a member of the IBM Watson Health Life Sciences Board; is a member of the Advisory Board for Element Science, the Advisory Board for Facebook, and the Physician Advisory Board for Aetna; and is the co-founder of Hugo Health, a personal health information platform, and co-founder of Refactor Health, a healthcare AI-augmented data management company, and Ensight-AI, Inc. Dr. Bhatt discloses the following relationships—Advisory Board: Angiowave, Bayer, Boehringer Ingelheim, CellProthera, Cereno Scientific, Elsevier Practice Update Cardiology, High Enroll, Janssen, Level Ex, McKinsey, Medscape Cardiology, Merck, MyoKardia, NirvaMed, Novo Nordisk, PhaseBio, PLx Pharma, Stasys; Board of Directors: American Heart Association New York City, Angiowave (stock options), Bristol Myers Squibb (stock), DRS.LINQ (stock options), High Enroll (stock); Consultant: Broadview Ventures, GlaxoSmithKline, Hims, SFJ, Youngene; Data Monitoring Committees: Acesion Pharma, Assistance Publique-Hôpitaux de Paris, Baim Institute for Clinical Research (formerly Harvard Clinical Research Institute, for the PORTICO trial, funded by St. Jude Medical, now Abbott), Boston Scientific (Chair, PEITHO trial), Cleveland Clinic, Contego Medical (Chair, PERFORMANCE 2), Duke Clinical Research Institute, Mayo Clinic, Mount Sinai School of Medicine (for the ENVISAGE trial, funded by Daiichi Sankyo; for the ABILITY-DM trial, funded by Concept Medical; for ALLAY-HF, funded by Alleviant Medical), Novartis, Population Health Research Institute; Rutgers University (for the NIH-funded MINT Trial); Honoraria: American College of Cardiology (Senior Associate Editor, Clinical Trials and News, ACC.org; Chair, ACC Accreditation Oversight Committee),

Arnold and Porter law firm (work related to Sanofi/Bristol-Myers Squibb clopidogrel litigation), Baim Institute for Clinical Research (formerly Harvard Clinical Research Institute; RE-DUAL PCI clinical trial steering committee funded by Boehringer Ingelheim; AEGIS-II executive committee funded by CSL Behring), Belvoir Publications (Editor in Chief, Harvard Heart Letter), Canadian Medical and Surgical Knowledge Translation Research Group (clinical trial steering committees), CSL Behring (AHA lecture), Cowen and Company, Duke Clinical Research Institute (clinical trial steering committees, including for the PRONOUNCE trial, funded by Ferring Pharmaceuticals), HMP Global (Editor in Chief, Journal of Invasive Cardiology), Journal of the American College of Cardiology (Guest Editor; Associate Editor), K2P (Co-Chair, interdisciplinary curriculum), Level Ex, Medtelligence/ReachMD (CME steering committees), MJH Life Sciences, Oakstone CME (Course Director, Comprehensive Review of Interventional Cardiology), Piper Sandler, Population Health Research Institute (for the COMPASS operations committee, publications committee, steering committee, and USA national co-leader, funded by Bayer), WebMD (CME steering committees), Wiley (steering committee); Other: Clinical Cardiology (Deputy Editor); Patent: Sotagliflozin (named on a patent for sotagliflozin assigned to Brigham and Women's Hospital who assigned to Lexicon; neither I nor Brigham and Women's Hospital receive any income from this patent); Research Funding: Abbott, Acesion Pharma, Afimmune, Aker Biomarine, Alnylam, Amarin, Amgen, AstraZeneca, Bayer, Beren, Boehringer Ingelheim, Boston Scientific, Bristol-Myers Squibb, Cardax, CellProthera, Cereno Scientific, Chiesi, CinCor, Cleerly, CSL Behring, Eisai, Ethicon, Faraday Pharmaceuticals, Ferring Pharmaceuticals, Forest Laboratories, Fractyl, Garmin, HLS Therapeutics, Idorsia, Ironwood, Ischemix, Janssen, Javelin, Lexicon, Lilly, Medtronic, Merck, Moderna, MyoKardia, NirvaMed, Novartis, Novo Nordisk, Otsuka, Owkin, Pfizer, PhaseBio, PLx Pharma, Recardio, Regeneron, Reid Hoffman Foundation, Roche, Sanofi, Stasys, Synaptic, The Medicines Company, Youngene, 89Bio; Royalties: Elsevier (Editor, Braunwald's Heart Disease); Site Co-Investigator: Abbott, Biotronik, Boston Scientific, CSI, Endotronix, St. Jude Medical (now Abbott), Philips, SpectraWAVE, Svelte, Vascular Solutions; Trustee: American College of Cardiology; Unfunded Research: FlowCo. All other authors declare no relevant competing interests.

Data availability

The development dataset from Yale and the validation datasets from Mount Sinai, Houston Methodist, and Lake Regional Hospital are not publicly available, given the stipulations of the relevant Institutional Review Boards. The external datasets from Brazil, UK Biobank, and PTB-XL are available directly from the respective groups and are outside the purview of the authors.

References

1. Schlant RC, Adolph RJ, DiMarco JP, Dreifus LS, Dunn MI, Fisch C, *et al.* Guidelines for electrocardiography. A report of the American College of Cardiology/American Heart Association task force on assessment of diagnostic and therapeutic cardiovascular procedures (committee on electrocardiography). *Circulation* 1992;**85**:1221–1228.
2. Shah AP, Rubin SA. Errors in the computerized electrocardiogram interpretation of cardiac rhythm. *J Electrocardiol* 2007;**40**:385–390.
3. Guglin ME, Thatai D. Common errors in computer electrocardiogram interpretation. *Int J Cardiol* 2006;**106**:232–237.
4. Schläpfer J, Wellens HJ. Computer-interpreted electrocardiograms: benefits and limitations. *J Am Coll Cardiol* 2017;**70**:1183–1192.
5. Garvey JL, Zegre-Hemsey J, Gregg R, Studnek JR. Electrocardiographic diagnosis of ST segment elevation myocardial infarction: an evaluation of three automated interpretation algorithms. *J Electrocardiol* 2016;**49**:728–732.
6. Estes NAM III. Computerized interpretation of ECGs: supplement not a substitute. *Circ Arrhythm Electrophysiol* 2013;**6**:2–4.
7. Bogun F, Anh D, Kalahasty G, Wissner E, Bou Serhal C, Bazzi R, *et al.* Misdiagnosis of atrial fibrillation and its clinical consequences. *Am J Med* 2004;**117**:636–642.
8. Jones CA, Park TS, Ahearn M, Mishra AK, Variyam JN. Health status and health care access of farm and rural populations. USDA Economic Research Service. 2009.
9. Aneja S, Ross JS, Wang Y, Matsumoto M, Rodgers GP, Bernheim SM, *et al.* US cardiologist workforce from 1995 to 2007: modest growth, lasting geographic maldistribution especially in rural areas. *Health Aff (Millwood)* 2011;**30**:2301–2309.
10. Soriano Marcolino M, Minelli Figueira R, Pereira Afonso Dos Santos J, Silva Cardoso C, Luiz Ribeiro A, Alkmim MB. The experience of a sustainable large scale Brazilian telehealth network. *Telemed J E Health* 2016;**22**:899–908.
11. Cook DA, Oh S-Y, Pusic MV. Accuracy of physicians' electrocardiogram interpretations. *JAMA Intern Med* 2020;**180**:1461.
12. Ribeiro AH, Ribeiro MH, Paixão GMM, Oliveira DM, Gomes PR, Canazart JA, *et al.* Automatic diagnosis of the 12-lead ECG using a deep neural network. *Nat Commun* 2020;**11**:1760.
13. Hannun AY, Rajpurkar P, Haghpanahi M, Tison GH, Bourn C, Turakhia MP, *et al.* Cardiologist-level arrhythmia detection and classification in ambulatory electrocardiograms using a deep neural network. *Nat Med* 2019;**25**:65–69.
14. Hughes JW, Olgin JE, Avram R, Abreau SA, Sittler T, Radia K, *et al.* Performance of a convolutional neural network and explainability technique for 12-lead electrocardiogram interpretation. *JAMA Cardiol* 2021;**6**:1285–1295.
15. Sangha V, Mortazavi BJ, Haimovich AD, Ribeiro AH, Brandt CA, Jacoby DL, *et al.* Automated multilabel diagnosis on electrocardiographic images and signals. *Nat Commun* 2022;**13**:1583.
16. Sangha V, Nargesi AA, Dhingra LS, Khunte A, Mortazavi BJ, Ribeiro AH, *et al.* Detection of left ventricular systolic dysfunction from electrocardiographic images. *Circulation* 2023;**148**:765–777.
17. Bao H, Dong L, Piao S, Wei F. BEiT: BERT Pre-Training of Image Transformers. *arXiv [cs.CV]*. https://doi.org/10.48550/arXiv.2106.08254, 15 June 2021; preprint:not peer reviewed.
18. Radford A, Wu J, Child R, Luan D, Amodei D, Sutskever I. Language Models are Unsupervised Multitask Learners. 2019. https://storage.prod.researchhub.com/uploads/papers/2020/06/01/language-models.pdf.
19. Lin C-Y. ROUGE: a package for automatic evaluation of summaries, In: *Text Summarization Branches out*. Barcelona, Spain: Association for Computational Linguistics; 2004. p74–81.
20. Papineni K, Roukos S, Ward T, Zhu W-J. BLEU: a method for automatic evaluation of machine translation. https://aclanthology.org/P02-1040.pdf (accessed June 13, 2023).
21. Banerjee S, Lavie A. METEOR: an automatic metric for MT evaluation with improved correlation with human judgments, In: *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*. Ann Arbor, Michigan: Association for Computational Linguistics; 2005. p65–72.
22. Vedantam R, Zitnick CL, Parikh D. CIDEr: Consensus-based image description evaluation. *arXiv [cs.CV]*. https://doi.org/10.48550/arXiv.1411.5726, 20 November 2014, preprint:not peer reviewed
23. Ribeiro ALP, Paixão GMM, Gomes PR, Ribeiro MH, Ribeiro AH, Canazart JA, *et al.* Tele-electrocardiography and bigdata: the CODE (clinical outcomes in digital electrocardiography) study. *J Electrocardiol* 2019;**57S**:S75–S78.
24. Selivanov A, Rogov OY, Chesakov D, Shelmanov A, Fedulova I, Dylov DV. Medical image captioning via generative pretrained transformers. *Sci Rep* 2023;**13**:4171.
25. Ayesha H, Iqbal S, Tariq M, Abrar M, Sanaullah M, Abbas I, *et al.* Automatic medical image interpretation: state of the art and future directions. *Pattern Recognit* 2021;**114**:107856.
26. Bartels MGG, Najdenkoska I, van de Leur RR, Sammani A, Taha K, Knigge DM, *et al.* Learning to automatically generate accurate ECG captions. *Proceedings of Machine Learning Research* 2022;**172**:86–102.
27. Kashou AH, Ko W-Y, Attia ZI, Cohen MS, Friedman PA, Noseworthy PA. A comprehensive artificial intelligence-enabled electrocardiogram interpretation program. *Cardiovasc Digit Health J* 2020;**1**:62–70.
28. Siontis KC, Noseworthy PA, Attia ZI, Friedman PA. Artificial intelligence-enhanced electrocardiography in cardiovascular disease management. *Nat Rev Cardiol* 2021;**18**:465–478.
29. Khunte A, Khera R, *et al.* ECG-GPT, CarDS Lab. 2024; cards-lab.org/ecg-gpt. 2025.