Supplementary Information
ECG-LLM Foundation Model for ECG-Based Cardiac Reasoning

Supplementary Information

1 Dataset statistics

Table 1 shows the number of ECG recordings and patients in each dataset split. Table 2 summarizes the final ECG-LLM instruction-tuning corpus.

Split	Entity	UK Biobank	MIMIC-IV-ECG	PTB-XL	EchoNext	All
Train	ECG studies	34,280	560,791	11,566	72,475	679,112
Train	Patients	34,280	115,508	10,403	26,218	186,409
Validation	ECG studies	–	62,251	1,274	4,626	68,151
Validation	Patients	–	12,834	1,155	4,626	18,615
Test	ECG studies	6,049	156,851	3,214	5,442	171,556
Test	Patients	6,049	32,086	2,891	5,442	46,468

Table 1: Dataset split statistics. The table reports the number of ECG studies and unique patients in the training, validation, and test splits for each ECG cohort. UK Biobank was split subject-wise into training and test sets only.

Dataset	QA source	Type	# Samples
UK Biobank	Structured-record QA	ECG and CMR, standard	670,545
	Structured-record QA	ECG and CMR, abnormal-enriched	173,238
	Structured-record QA	ECG and CMR, chain-of-thought	167,701
EchoNext	Structured-record QA	ECG and echo, standard	546,156
	Structured-record QA	ECG and echo, chain-of-thought	129,625
MIMIC-IV-ECG	Structured-record QA	ECG interpretation, standard	2,775,751
	ECG-QA	Single-Verify	125,380
	ECG-QA	Single-Choose	107,860
	ECG-QA	Single-Query	153,156
	ECGInstruct	Feature (Close/Open)	3,865
	ECGInstruct	Rhythm (Close/Open)	38,124
	ECGInstruct	Morphology (Close/Open)	62,663
	ECGInstruct	Report (Open)	345,561
PTB-XL	Structured-record QA	ECG interpretation, standard	127,219
	ECG-QA	Single-Verify	62,554
	ECG-QA	Single-Choose	50,015
	ECG-QA	Single-Query	46,737
	ECGInstruct	Feature (Close/Open)	57,108
	ECGInstruct	Rhythm (Close/Open)	18,899
	ECGInstruct	Morphology (Close/Open)	108,379
	ECGInstruct	Report (Open)	7,890
Total			**5,778,426**

Table 2: Final ECG-LLM instruction-tuning corpus. Structured-record QA denotes generated examples produced from the per-study cardiac records created in this work.

2 UK Biobank ECG and CMR phenotype question bank

Question group	ECG-LLM questions	Target phenotypes
Heart rate and rhythm	What does the ECG reveal about the patient's heart rate and rhythm? Review the ECG and describe the rhythm and rate.	ECG heart rate [bpm]; RR interval [ms].
P-wave characteristics and atrioventricular intervals	Review this ECG and describe atrial activity and AV conduction. Review the ECG and infer the P-wave characteristics and atrioventricular intervals.	P duration [ms]; PP interval [ms]; PQ interval [ms].
QRS duration and ventricular conduction	Is the QRS duration within normal limits? Review the ECG and identify any conduction delays.	QRS duration [ms].
QT interval and ventricular repolarization	Are there any repolarization abnormalities on this ECG? Is the QT interval within normal limits?	QT interval [ms]; QTC interval [ms].
Electrical axes	Are the electrical axes within normal limits? Does the ECG show any axis deviation?	P axis [degrees]; R axis [degrees]; T axis [degrees].
Left ventricular systolic function	Review the ECG and describe LV ejection fraction. What does this ECG reveal about left ventricular systolic function?	LV ejection fraction.
Left ventricular stroke volume	Review the ECG and describe LV stroke volume. What does this ECG reveal about left ventricular stroke volume?	LV stroke volume.
Left ventricular volumes	Review the ECG and describe any abnormalities found for LV. What does this ECG reveal about left ventricular volumes?	LV end diastolic volume; LV end systolic volume.
Left ventricular myocardial mass and wall thickness	Is there any evidence of ventricular hypertrophy or increased myocardial mass? What does this ECG reveal about left ventricular wall thickness?	LV myocardial mass; LV mean myocardial wall thickness global.
Left ventricular strain	Do the strain measurements suggest any myocardial dysfunction? Review this ECG and describe any abnormalities in left ventricular strain.	LV circumferential strain global; LV longitudinal strain global; LV radial strain global.
Cardiac output and cardiac index	What does this ECG indicate about cardiac output and cardiac index? Examine this ECG and tell if the cardiac output or index within normal limits?	LV cardiac output; cardiac index [litres/min/m2].
Right ventricular stroke volume	Review this ECG and describe RV stroke volume. What does this ECG reveal about right ventricular stroke volume?	RV stroke volume.
Right ventricular ejection fraction	Review this ECG and describe RV ejection fraction. What does this ECG reveal about right ventricular systolic function?	RV ejection fraction.
Right ventricular volumes and function	Review this ECG and describe all abnormalities for RV. Review the ECG and indicate about abnormalities in right ventricular volumes.	RV end diastolic volume; RV end systolic volume.
Left atrial volumes and function	What does this ECG suggest about left atrial volumes and function? Examine this ECG and tell if there is any evidence of left atrial enlargement or dysfunction?	LA maximum volume; LA minimum volume; LA stroke volume; LA ejection fraction.
Right atrial volumes and function	What does this ECG reveal about right atrial volumes and function? Examine this ECG and tell if there is any evidence of right atrial enlargement or dysfunction?	RA maximum volume; RA minimum volume; RA stroke volume; RA ejection fraction.
Aortic distensibility	Review this ECG and state whether there is any indirect evidence that could suggest abnormal aortic or vascular stiffness.	Ascending aorta distensibility; Descending aorta distensibility.

Table 3: UK Biobank question bank used for ECG-derived and CMR-derived phenotype elicitation from ECG alone.

3 UK Biobank phenotype extraction protocol

Because ECG-LLM was trained to generate free-text answers, the UK Biobank phenotype evaluation required an additional mapping step from generated text to the predefined phenotype-category schema. We used the Google LangExtract framework [1, 2] to apply a schema-constrained extraction prompt to ECG-LLM answers. LangExtract is a Python library for extracting structured information from unstructured text using LLMs and user-defined extraction instructions. In our setup, the backend model was a locally deployed GPT-OSS-20B instance [3, 4, 5].

For each ECG and question group, the extractor received the generated answer, the target phenotype fields assigned to that question group, and the allowed category labels for each target phenotype. It was instructed to return allowed category labels only for those target fields found in the ECG-LLM answers. This group-specific extraction kept the evaluation aligned with the clinical intent of the question: the model was neither rewarded nor penalized for mentioning unrelated phenotypes outside the asked group. This step was necessary because a free-text answer could describe several related phenotypes together, refer to multiple measurements collectively, or express a category without repeating the exact phenotype name used in the evaluation schema. Therefore, simple string matching or regular expressions would not reliably map answers to field-specific category labels. The extractor served only as a mapping layer from free text to the predefined category space used for evaluation.

Prompt for schema-constrained phenotype extraction

```
Extract discrete cardiac phenotype categories.
Use only the phenotype field names listed below as possible extraction classes.
Use the allowed category words as the extraction text.
Do not extract numeric values or explanations.
If the answer says several named phenotypes are normal, extract normal for each named
↪  phenotype.
Sometimes, phenotype fields may not contain dimension information in square brackets, this is
↪  still valid.
If the answer says one phenotype is abnormal but does not give a discrete allowed category,
↪  do not extract it.
ALLOWED PHENOTYPES AND CATEGORIES:
- <phenotype field 1>: <allowed category 1>, <allowed category 2>, ...
- <phenotype field 2>: <allowed category 1>, <allowed category 2>, ...
- ...
```

4 UK Biobank phenotype extraction results

Field	n	K	Cov.	Miss.	Acc.	BAcc.	wF1	Maj.	mF1
ECG measurements									
ECG heart rate [bpm]	1000	5	1.000	0.000	0.958	0.974	0.958	0.126	0.977
R axis [degrees]	1000	3	1.000	0.000	0.987	0.919	0.987	0.315	0.940
RR interval [ms]	1000	5	1.000	0.000	0.944	0.969	0.944	0.127	0.926
P duration [ms]	979	2	1.000	0.000	0.974	0.824	0.972	0.481	0.886
QRS duration [ms]	1000	3	0.987	0.013	0.976	0.886	0.976	0.320	0.874
T axis [degrees]	1000	4	1.000	0.000	0.888	0.681	0.880	0.212	0.726
QTC interval [ms]	1000	3	1.000	0.000	0.962	0.652	0.958	0.324	0.719
PP interval [ms]	999	5	1.000	0.000	0.930	0.605	0.927	0.126	0.637
P axis [degrees]	938	3	1.000	0.000	0.908	0.540	0.888	0.310	0.614
PQ interval [ms]	980	3	1.000	0.000	0.906	0.458	0.879	0.313	0.522
CMR-derived phenotypes									
LV mean myocardial wall thickness global	990	2	1.000	0.000	0.965	0.974	0.966	0.445	0.947
RV end diastolic volume	1000	3	0.999	0.001	0.993	0.708	0.991	0.327	0.708
LV radial strain global	833	3	0.975	0.025	0.998	0.667	0.996	0.311	0.666
RV end systolic volume	1000	3	0.999	0.001	0.995	0.667	0.993	0.327	0.641
RA minimum volume	988	3	1.000	0.000	0.910	0.594	0.885	0.314	0.592
RA maximum volume	989	3	0.999	0.001	0.906	0.569	0.894	0.312	0.560
LA maximum volume	989	3	0.999	0.001	0.950	0.473	0.937	0.322	0.547
LV end systolic volume	998	3	1.000	0.000	0.943	0.460	0.930	0.318	0.507
LV myocardial mass	1000	2	0.977	0.023	0.950	0.495	0.934	0.488	0.487
Ascending aorta distensibility	872	2	0.814	0.186	0.866	0.484	0.891	0.489	0.486
RV ejection fraction	999	5	1.000	0.000	0.981	0.443	0.974	0.197	0.469
LA minimum volume	988	3	1.000	0.000	0.881	0.429	0.850	0.305	0.459
RV stroke volume	999	3	1.000	0.000	0.961	0.472	0.955	0.327	0.458
RA stroke volume	989	3	0.999	0.001	0.884	0.403	0.866	0.314	0.416
LV end diastolic volume	1000	3	0.998	0.002	0.945	0.383	0.925	0.322	0.410
Cardiac index [litres/min/m^2]	843	5	1.000	0.000	0.696	0.396	0.672	0.114	0.397
LA ejection fraction	989	3	0.999	0.001	0.908	0.369	0.868	0.316	0.381
LV ejection fraction	1000	5	1.000	0.000	0.946	0.321	0.934	0.191	0.371
LV stroke volume	1000	3	1.000	0.000	0.941	0.360	0.917	0.323	0.371
RA ejection fraction	989	3	0.999	0.001	0.876	0.361	0.831	0.310	0.363
LV longitudinal strain global	815	3	0.996	0.004	0.966	0.346	0.950	0.327	0.352
Descending aorta distensibility	872	3	0.810	0.190	0.857	0.358	0.887	0.326	0.339
LA stroke volume	989	3	0.999	0.001	0.956	0.333	0.935	0.326	0.326
LV circumferential strain global	833	3	0.975	0.025	0.950	0.333	0.925	0.324	0.325
LV cardiac output	1000	5	0.843	0.157	0.766	0.271	0.695	0.170	0.277

Table 4: Full UK Biobank phenotype extraction results for ECG-derived and CMR-derived fields. The table reports field-level sample size (n), number of possible classification classes (K), coverage, missing rate, accuracy, balanced accuracy, weighted-F1, majority-class macro-F1, and macro-F1. Fields are ordered by macro-F1 within each group.

5 EchoNext ECG-only binary phenotype questions

Phenotype	ECG-LLM question
LVEF $\leq$ 45%	Does this ECG suggest moderately or severely reduced left ventricular systolic function?
LVWT $\geq$ 1.3 cm	Does this ECG show moderate left ventricular hypertrophy with the maximum of the interventricular septum (IVS) or posterior wall (LVPW) thickness greater than or equal to 1.3 cm?
AS $\geq$ moderate	Does this ECG show moderate or severe aortic stenosis?
AR $\geq$ moderate	Does this ECG suggest moderate or severe aortic regurgitation?
MR $\geq$ moderate	Does this ECG suggest moderate or severe mitral regurgitation?
TR $\geq$ moderate	Does this ECG suggest moderate or severe tricuspid regurgitation?
PR $\geq$ moderate	Does this ECG suggest moderate or severe pulmonary regurgitation?
RV systolic dysf. $\geq$ moderate	Does this ECG suggest moderate or severe right ventricular systolic dysfunction?
Pericardial eff. $\geq$ moderate/large	Does this ECG suggest a moderate or large pericardial effusion?
PASP $\geq$ 45 mmHg	Does this ECG suggest pulmonary hypertension, as would be expected with a pulmonary artery systolic pressure of at least 45 mmHg?
TR Vmax $\geq$ 3.2 m/s	Does this ECG suggest a tricuspid regurgitation jet velocity of at least 3.2 m/s?

Table 5: EchoNext phenotype-specific questions used for ECG-only yes/no question answering. For each ECG and phenotype, ECG-LLM first answered the initial question in free text. In the same conversation, the model was then asked: "Give the final answer to the original question as yes or no." The final yes/no response was parsed and compared with the corresponding echocardiographic binary label.

6 PTB-XL ECG-QA performance by question format and attribute

Question format	Attribute type	n	EM	μP	μR	μF1
Single-Verify	SCP code	8,243	79.52	79.52	79.52	79.52
Single-Verify	Heart axis	237	89.87	89.87	89.87	89.87
Single-Verify	Stage of infarction	168	82.74	82.74	82.74	82.74
Single-Verify	Noise	3,183	67.73	67.73	67.73	67.73
Single-Verify	Extra systole	230	77.39	77.39	77.39	77.39
Single-Verify	Numeric feature	1,020	77.84	77.84	77.84	77.84
Single-Choose	SCP code	9,358	69.30	77.50	72.95	75.16
Single-Choose	Heart axis	36	75.00	75.00	75.00	75.00
Single-Choose	Stage of infarction	18	55.56	55.56	55.56	55.56
Single-Choose	Noise	317	43.22	55.21	46.27	50.35
Single-Choose	Extra systole	22	50.00	54.55	46.15	50.00
Single-Choose	Numeric feature	104	60.58	60.58	60.58	60.58
Single-Query	SCP code	9,208	43.07	59.47	57.99	58.72
Single-Query	Heart axis	167	74.85	74.85	74.85	74.85
Single-Query	Stage of infarction	178	46.07	46.07	46.07	46.07
Single-Query	Noise	3,960	38.86	44.48	46.04	45.24
Single-Query	Extra systole	241	54.77	56.85	55.02	55.92
Single-Query	Numeric feature	4,403	25.03	65.75	65.96	65.85

Table 6: PTB-XL ECG-QA performance stratified by question format and ECG-QA attribute type. Question format is the single-ECG ECG-QA format evaluated in this work. Attribute type is the broad ECG-QA metadata category used to generate the question. n is the number of evaluated question instances in the corresponding question-format and attribute-type subset. EM is exact-match accuracy over the complete predicted answer set. μP, μR, and μF1 are micro-averaged option-level precision, recall, and F1.

7 MIMIC-IV-ECG ECG-QA performance by question format and attribute

Question format	Attribute type	n	EM	μP	μR	μF1
Single-Verify	SCP code	23,950	81.16	81.16	81.16	81.16
Single-Verify	Heart axis	240	81.67	81.67	81.67	81.67
Single-Verify	Stage of infarction	180	78.33	78.33	78.33	78.33
Single-Verify	Noise	148	79.05	79.05	79.05	79.05
Single-Verify	Numeric feature	1,020	79.22	79.22	79.22	79.22
Single-Choose	SCP code	21,971	68.43	75.41	71.82	73.57
Single-Choose	Heart axis	38	50.00	55.26	52.50	53.85
Single-Choose	Stage of infarction	18	77.78	77.78	77.78	77.78
Single-Choose	Numeric feature	104	60.58	60.58	60.58	60.58
Single-Query	SCP code	54,167	19.25	55.34	48.35	51.61
Single-Query	Heart axis	200	64.50	64.50	64.50	64.50
Single-Query	Stage of infarction	250	44.80	44.80	44.80	44.80
Single-Query	Noise	27	7.41	12.50	12.90	12.70
Single-Query	Numeric feature	7,617	20.90	69.54	66.92	68.21

Table 7: MIMIC-IV-ECG ECG-QA performance stratified by question format and ECG-QA attribute type. The question format is the single-ECG ECG-QA format evaluated in this work. Attribute type is the broad ECG-QA metadata category used to generate the question. n is the number of evaluated question instances in the corresponding question-format and attribute-type subset. EM is exact-match accuracy over the complete predicted answer set. μP, μR, and μF1 are micro-averaged option-level precision, recall, and F1.

8 ECG-QA Single-Verify performance by attribute

Table 8: ECG-QA Single-Verify performance by attribute for PTB-XL and MIMIC-IV-ECG. EM is reported as percent.

Attribute	MIMIC		PTB-XL	
	n	EM	n	EM
SCP-code diagnostic statements				
2:1 AV block	60	93.3	–	–
2:1 sinoatrial block	42	92.9	–	–
3:1 AV block	60	70.0	–	–
4:1 AV block	60	73.3	–	–
aberrant supraventricular complexes	60	75.0	–	–
aberrant ventricular complex	60	90.0	–	–
abnormal qrs	–	–	60	71.7
abnormal R wave progression	60	83.3	–	–
abnormal ventricular conduction pathways	60	73.3	–	–
accelerated idioventricular rhythm	60	88.3	–	–
accelerated junctional rhythm	80	65.0	–	–
any diagnostic symptoms	120	75.8	120	80.8
any form-related symptoms	60	80.0	60	66.7
any kind of abnormal symptoms	60	81.7	60	71.7
any rhythm-related symptoms	60	75.0	60	65.0
atrial arrhythmia	12	100.0	–	–
atrial bigeminy	60	66.7	–	–
atrial couplet	60	65.0	–	–
atrial fibrillation	60	90.0	60	86.7
atrial flutter	80	63.7	36	94.4
atrial premature complex	–	–	60	71.7
atrial tachycardia	60	91.7	–	–
atypical left bundle branch block	60	91.7	–	–
atypical right bundle branch block	60	85.0	–	–
AV dissociation	60	76.7	–	–
AV sequential pacemaker	60	90.0	–	–
bi-atrial overload/enlargement	80	68.8	–	–
bigeminal pattern (unknown origin, supraventricular, or ventricular)	–	–	50	80.0
Biventricular hypertrophy	60	91.7	–	–
Broad R wave in lateral leads	60	86.7	–	–
complete left bundle branch block	60	93.3	60	100.0
complete right bundle branch block	60	98.3	61	96.7
conduction disturbance	60	71.7	60	83.3
Deep S wave	581	91.7	–	–
dextrocardia	56	78.6	–	–
digitalis effect	–	–	67	76.1
dual chamber electronic pacing	60	98.3	–	–
early R wave transition	60	95.0	–	–
early repolarization	7	85.7	–	–
ectopic atrial bradycardia	60	88.3	–	–
ectopic atrial rhythm	60	73.3	–	–
ectopic atrial tachycardia	60	91.7	–	–
electronic atrial pacing	60	83.3	–	–
extreme tachycardia	59	93.2	–	–
first degree AV block	60	83.3	63	66.7
fusion complexes	60	86.7	–	–
high amplitude T-waves	60	91.7	–	–
high grade AV block	58	77.6	–	–
high P-voltages	426	82.4	–	–
high QRS voltage	600	86.8	145	86.2
hypertrophy	60	83.3	60	88.3
idioventricular rhythm	60	86.7	–	–
incomplete left bundle branch block	60	91.7	51	92.2
incomplete right bundle branch block	60	88.3	69	71.0
intermittent second degree AV block	60	80.0	–	–
intraventricular conduction disturbance	60	88.3	–	–
inverted T-waves	60	70.0	621	82.0
ischemic in anterior leads	–	–	32	84.4
ischemic in anterolateral leads	–	–	80	61.3
ischemic in anteroseptal leads	–	–	63	84.1
ischemic in inferior leads	–	–	66	74.2
ischemic in inferolateral leads	–	–	67	62.7
ischemic in lateral leads	–	–	55	65.5
ischemic ST-T changes	60	85.0	–	–
ischemic ST-T changes in anterior leads	80	66.2	–	–
ischemic ST-T changes in anteroseptal leads	80	68.8	–	–
ischemic ST-T changes in diffuse leads	60	65.0	–	–
ischemic ST-T changes in inferior leads	80	62.5	–	–
ischemic ST-T changes in inferoseptal leads	80	77.5	–	–
ischemic ST-T changes in lateral leads	14	71.4	–	–
ischemic ST-T changes in septal leads	80	76.2	–	–
junctional bradycardia	50	80.0	–	–
junctional rhythm	80	67.5	–	–

Attribute	MIMIC		PTB-XL	
	n	EM	n	EM
junctional tachycardia	60	93.3	–	–
late R wave transition	30	86.7	–	–
left anterior fascicular block	80	77.5	68	82.4
left atrial overload/enlargement	80	43.8	80	55.0
left posterior fascicular block	60	85.0	60	98.3
Left ventricular hypertrophy	80	65.0	80	78.8
long QT interval	60	88.3	52	78.8
low amplitude t-wave	–	–	636	78.0
low QRS voltage	1260	93.7	–	–
low qrs voltages in the frontal and horizontal leads	–	–	126	88.1
low R	438	75.3	–	–
Mobitz type 1 second-degree AV block	60	73.3	–	–
Mobitz type 2 second-degree AV block	49	83.7	–	–
multifocal premature ventricular complexes	60	88.3	–	–
Myocardial infarction	60	85.0	60	78.3
Myocardial infarction in anterior leads	80	71.2	80	67.5
Myocardial infarction in anterolateral leads	80	63.7	76	78.9
Myocardial infarction in anteroseptal leads	80	57.5	80	66.2
Myocardial infarction in inferior leads	80	67.5	80	75.0
Myocardial infarction in inferolateral leads	80	73.8	79	68.4
myocardial infarction in inferoposterior leads	–	–	7	71.4
myocardial infarction in inferoposterolateral leads	–	–	7	85.7
Myocardial infarction in lateral leads	80	66.2	49	81.6
Myocardial infarction in posterior leads	80	75.0	7	85.7
Myocardial infarction in septal leads	80	72.5	–	–
non-diagnostic T abnormalities	1260	87.2	743	79.3
non-specific intraventricular conduction disturbance (block)	–	–	63	76.2
non-specific ischemic	–	–	80	70.0
non-specific ST changes	1080	77.6	456	77.4
non-specific ST depression	198	85.9	656	80.5
non-specific st elevation	–	–	24	83.3
non-specific t abnormality	1260	78.4	–	–
non-specific T-wave changes	1080	77.9	684	75.4
non-sustained ventricular tachycardia	60	93.3	–	–
Normal ECG	60	85.0	60	86.7
normal functioning artificial pacemaker	–	–	48	100.0
notched P wave	49	83.7	–	–
P wave abnormality	60	73.3	–	–
pacemaker activity	60	85.0	–	–
pacemaker rhythm	60	96.7	–	–
paired ventricular premature complexes	60	80.0	–	–
paroxysmal idioventricular rhythm	60	73.3	–	–
poor R wave progression	420	76.0	–	–
premature atrial complexes	60	73.3	–	–
premature ventricular complexes	60	85.0	–	–
premature ventricular interpolated complexes	60	83.3	–	–
Prolonged PR interval	60	68.3	60	81.7
prolonged QRS duration	60	85.0	–	–
Q waves present	1039	80.2	647	83.2
QRS changes in anteroseptal leads	60	76.7	–	–
QTc prolongation	60	73.3	–	–
rapid ventricular response	60	88.3	–	–
regular rhythm	60	76.7	–	–
repolarization abnormality	1260	81.0	–	–
Reversed R wave progression	52	82.7	–	–
right atrial overload/enlargement	60	85.0	58	91.4
right ventricular conduction delay	59	71.2	–	–
Right ventricular hypertrophy	80	70.0	–	–
rSr' type in V1 or V2	60	95.0	–	–
S1 S2 S3 type QRS pattern	36	91.7	–	–
second degree AV block	60	78.3	–	–
second-degree SA block type II	60	68.3	–	–
Short PR interval	60	70.0	–	–
short QT interval	60	78.3	–	–
short QTc	30	86.7	–	–
Significant repolarization change	60	73.3	–	–
sinus arrhythmia	60	60.0	60	71.7
sinus bradycardia	60	95.0	60	83.3
sinus pause	60	71.7	–	–
sinus rhythm	60	80.0	60	76.7
sinus tachycardia	80	92.5	60	98.3
slow ventricular response	60	91.7	–	–
ST Depression	1260	84.0	–	–
ST Elevation	1020	77.3	–	–
ST(-T) change	1260	79.1	–	–
st/t change	–	–	60	78.3
subendocardial injury in anterolateral leads	–	–	61	88.5
subendocardial injury in anteroseptal leads	–	–	63	88.9
subendocardial injury in inferolateral leads	–	–	6	83.3
subendocardial injury in lateral leads	–	–	13	84.6
supraventricular bigeminy BIGU bigeminal pattern	60	60.0	–	–

Attribute	MIMIC		PTB-XL	
	n	EM	n	EM
supraventricular extrasystoles	60	61.7	–	–
supraventricular premature complex	60	88.3	–	–
supraventricular rhythm	60	71.7	–	–
supraventricular tachycardia	80	76.2	18	77.8
T-wave abnormality	1260	79.9	150	80.7
tall R wave in V1 or V2	60	93.3	–	–
tall R wave in V5 or V6	57	86.0	–	–
third degree AV block	60	81.7	–	–
TU fusion	52	78.8	–	–
unclassified aberrantly conducted complexes	60	78.3	–	–
uncontrolled ventricular response	60	98.3	–	–
ventricular bigeminy	60	86.7	–	–
ventricular couplet	60	80.0	–	–
ventricular escape rhythm	24	79.2	–	–
ventricular premature complex	–	–	60	88.3
ventricular trigeminy	60	95.0	–	–
ventricular-paced complexes or rhythm	60	95.0	–	–
Voltage criteria (QRS) for left ventricular hypertrophy	60	88.3	180	88.3
wandering pacemaker	36	77.8	–	–
wide QRS tachycardia	60	95.0	–	–
Wolff-Parkinson type A	6	83.3	–	–
Wolff-Parkinson type B	50	82.0	–	–
Wolff-Parkinson-White syndrome	60	85.0	–	–
Heart axis				
extreme axis deviation	60	90.0	57	93.0
left axis deviation	60	66.7	60	80.0
normal heart axis	60	76.7	60	93.3
right axis deviation	60	93.3	60	93.3
Stage of infarction				
early stage of myocardial infarction	60	83.3	59	89.8
middle stage of myocardial infarction	60	80.0	60	80.0
old stage of myocardial infarction	60	71.7	49	77.6
Noise				
any kind of noises	–	–	780	65.9
baseline drift	–	–	780	66.7
baseline wander	148	79.1	–	–
burst noise	–	–	777	68.1
electrodes problems	–	–	66	81.8
static noise	–	–	780	69.1
Numeric features				
above the normal range of p duration	60	68.3	60	68.3
above the normal range of pr interval	60	81.7	60	83.3
above the normal range of qrs duration	60	78.3	60	78.3
above the normal range of qt corrected	60	70.0	60	75.0
above the normal range of qt interval	60	81.7	60	70.0
above the normal range of rr interval	60	91.7	60	98.3
below the normal range of pr interval	60	66.7	60	61.7
below the normal range of qrs duration	60	85.0	60	85.0
below the normal range of qt corrected	60	78.3	60	81.7
below the normal range of qt interval	60	81.7	60	83.3
below the normal range of rr interval	60	100.0	60	93.3
within the normal range of p duration	60	73.3	60	65.0
within the normal range of pr interval	60	68.3	60	71.7
within the normal range of qrs duration	60	78.3	60	63.3
within the normal range of qt corrected	60	70.0	60	73.3
within the normal range of qt interval	60	73.3	60	76.7
within the normal range of rr interval	60	100.0	60	95.0
Extra systole				
any kind of extra systoles	–	–	50	82.0
extrasystoles	–	–	60	81.7
supraventricular extrasystoles	–	–	60	66.7
ventricular extrasystoles	–	–	60	80.0

9 Fields used for structured-record QA generation

Table 9: Dataset-specific fields used to construct structured per-study records for synthetic QA generation.

Field	Dataset	Comment
Patient context		
Patient age	UK Biobank, MIMIC-IV-ECG, EchoNext, PTB-XL	MIMIC-IV-ECG uses approximate anchor age.
Patient sex	UK Biobank, MIMIC-IV-ECG, EchoNext, PTB-XL	–
BMI	UK Biobank	–
Height	UK Biobank	–
Body surface area	UK Biobank	Used for indexed CMR reference ranges.
Chest pain or discomfort	UK Biobank	–
Shortness of breath on level ground	UK Biobank	–
Race or ethnicity	EchoNext	–
ECG measurements		
ECG heart rate	MIMIC-IV-ECG, UK Biobank	In MIMIC-IV-ECG, derived from RR interval if heart rate is unavailable.
Ventricular rate	EchoNext	–
Atrial rate	EchoNext	–
RR interval	MIMIC-IV-ECG, UK Biobank	–
PP interval	MIMIC-IV-ECG, UK Biobank	–
PR interval	MIMIC-IV-ECG, EchoNext	–
PQ interval	MIMIC-IV-ECG, UK Biobank	Derived from P onset and Q onset when needed.
P duration	MIMIC-IV-ECG, UK Biobank	Derived from P onset and P offset when needed.
QRS duration	MIMIC-IV-ECG, UK Biobank, EchoNext	In MIMIC-IV-ECG, derived from QRS onset and QRS end when needed.
QT interval	MIMIC-IV-ECG, UK Biobank	Used to derive QTc when needed.
QTc interval	MIMIC-IV-ECG, UK Biobank, EchoNext	In MIMIC-IV-ECG and UK Biobank, derived from QT and RR when needed.
P axis	MIMIC-IV-ECG, UK Biobank	–
R axis	MIMIC-IV-ECG, UK Biobank, PTB-XL	In PTB-XL, heart-axis labels are mapped to unified R-axis categories. MID: normal; LAD and ALAD: leftward; RAD and ARAD: rightward.
T axis	MIMIC-IV-ECG, UK Biobank	–
ECG report-derived fields		
ECG diagnosis and machine-report diagnosis text	MIMIC-IV-ECG, UK Biobank	ECG diagnosis is retained as text. For MIMIC-IV-ECG, diagnosis text is built from machine-measurement report columns.
German ECG report	PTB-XL	Original PTB-XL report text.
English ECG report	PTB-XL	English translation of the PTB-XL report generated using GPT-OSS-20B.
SCP-ECG diagnostic codes	PTB-XL	SCP codes are decoded into textual diagnostic descriptions.
Infarction stage	PTB-XL	–
CMR-derived fields		
LV end diastolic volume	UK Biobank	–
LV end systolic volume	UK Biobank	–
LV stroke volume	UK Biobank	–
LV myocardial mass	UK Biobank	–
LV ejection fraction	UK Biobank	–
LV cardiac output	UK Biobank	–
Cardiac index	UK Biobank	–
LV mean myocardial wall thickness global	UK Biobank	–
LV circumferential strain global	UK Biobank	–
LV longitudinal strain global	UK Biobank	–
LV radial strain global	UK Biobank	–
RV end diastolic volume	UK Biobank	–

Field	Dataset	Comment
RV end systolic volume	UK Biobank	–
RV stroke volume	UK Biobank	–
RV ejection fraction	UK Biobank	–
LA maximum volume	UK Biobank	–
LA minimum volume	UK Biobank	–
LA stroke volume	UK Biobank	–
LA ejection fraction	UK Biobank	–
RA maximum volume	UK Biobank	–
RA minimum volume	UK Biobank	–
RA stroke volume	UK Biobank	–
RA ejection fraction	UK Biobank	–
Ascending aorta distensibility	UK Biobank	–
Descending aorta distensibility	UK Biobank	–
EchoNext echocardiographic categorical fields		
Aortic stenosis	EchoNext	Severity category.
Aortic regurgitation	EchoNext	Severity category.
Mitral regurgitation	EchoNext	Severity category.
Tricuspid regurgitation	EchoNext	Severity category.
Pulmonary regurgitation	EchoNext	Severity category.
Right ventricular systolic function	EchoNext	Qualitative systolic function category.
Pericardial effusion	EchoNext	Presence and size category.
EchoNext echocardiographic numeric fields		
Interventricular septal thickness	EchoNext	–
Left ventricular posterior wall thickness	EchoNext	–
Left ventricular systolic function	EchoNext	Derived from LVEF value in percent.
EchoNext binary structural heart disease fields		
Left ventricular ejection fraction is 45% or lower	EchoNext	Binary descriptor derived from LVEF.
Left ventricular wall thickness is at least 1.3 cm	EchoNext	Binary descriptor derived from LV wall thickness.
Aortic stenosis is moderate or greater	EchoNext	Binary descriptor derived from aortic stenosis severity.
Aortic regurgitation is moderate or greater	EchoNext	Binary descriptor derived from aortic regurgitation severity.
Mitral regurgitation is moderate or greater	EchoNext	Binary descriptor derived from mitral regurgitation severity.
Tricuspid regurgitation is moderate or greater	EchoNext	Binary descriptor derived from tricuspid regurgitation severity.
Pulmonary regurgitation is moderate or greater	EchoNext	Binary descriptor derived from pulmonary regurgitation severity.
Right ventricular systolic dysfunction is moderate or greater	EchoNext	Binary descriptor derived from right ventricular systolic function.
Pericardial effusion is moderate or large	EchoNext	Binary descriptor derived from pericardial effusion size.
Pulmonary artery systolic pressure is at least 45 mmHg	EchoNext	Binary descriptor derived from PASP.
Tricuspid regurgitation max velocity is at least 3.2 m/s	EchoNext	Binary descriptor derived from maximum TR jet velocity.

10 Clinical categorization rules

Table 10: Clinical categorisation rules used to convert numeric measurements into textual descriptors before synthetic QA generation.

Measurement	Categorisation rule
ECG measurements	
ECG heart rate [bpm], ventricular rate, atrial rate	≤ 50: marked bradycardia; > 50 to ≤ 60: bradycardia; > 60 to ≤ 100: normal; > 100 to ≤ 120: mild tachycardia; > 120: marked tachycardia.
RR interval [ms], PP interval [ms]	Convert interval to rate as 60000/interval. Rate > 120: markedly short; > 100 to ≤ 120: short; ≥ 60 to ≤ 100: normal; ≥ 50 to < 60: prolonged; < 50: markedly prolonged.
P duration [ms]	If missing, derive as P offset − P onset. Then categorise as: < 120: normal; ≥ 120: prolonged.
PQ interval [ms], PR interval [ms]	If PQ is missing, derive as Q onset − P onset. Then categorise as: < 120: short; 120–200: normal; > 200: prolonged.
QRS duration [ms]	< 110: normal; ≥ 110 to < 120: mildly prolonged; ≥ 120: prolonged.
QTC interval [ms]	If QTC is unavailable, derive it from QT and RR using Bazett correction: $\mathrm{QTc} = \mathrm{QT}/\sqrt{\mathrm{RR}/1000}$. Female: ≤ 470: normal; > 470 to ≤ 490: borderline; > 490: prolonged. Male or unavailable sex: ≤ 450: normal; > 450 to ≤ 480: borderline; > 480: prolonged.
P axis [degrees]	< 0: leftward; 0–75: normal; > 75: rightward.
R axis [degrees]	< -30: leftward; −30–90: normal; > 90: rightward.
T axis [degrees]	< -15: leftward; -15 to < 15: borderline; 15–75: normal; > 75 to ≤ 105: borderline; > 105: rightward.
Ventricular function and cardiac output	
LV ejection fraction, left ventricular systolic function	Female: < 30: severely reduced; ≥ 30 to < 40: moderately reduced; ≥ 40 to < 52: mildly reduced; 52–79: normal; > 79: hyperdynamic. Male or unavailable sex: < 30: severely reduced; ≥ 30 to < 40: moderately reduced; ≥ 40 to < 49: mildly reduced; 49–79: normal; > 79: hyperdynamic.
RV ejection fraction	Female: < 30: severely reduced; ≥ 30 to < 40: moderately reduced; ≥ 40 to < 46: mildly reduced; 46–74: normal; > 74: hyperdynamic. Male or unavailable sex: < 30: severely reduced; ≥ 30 to < 40: moderately reduced; ≥ 40 to < 42: mildly reduced; 42–72: normal; > 72: hyperdynamic.
cardiac index [litres/min/m2]	< 2.2: low output; 2.2 to ≤ 2.5: borderline low; > 2.5 to ≤ 4.0: normal; > 4.0 to ≤ 4.5: borderline high; > 4.5: high output.
LV cardiac output	If body surface area is available, first compute cardiac index and apply the cardiac-index rule. Otherwise, use absolute ranges. Male 3.4–7.8: normal; female 2.7–6.3: normal; unknown sex 2.7–7.8: normal. Below range: low output; above range: high output.
LV mean myocardial wall thickness global	< 5.0: thinned; 5.0–11.0: normal; > 11.0: hypertrophied.
Echocardiographic LV wall thickness [cm], interventricular septal thickness, left ventricular posterior wall thickness	< 0.6: thinned. Male: ≤ 1.0: normal; > 1.0 to ≤ 1.3: mildly increased; > 1.3 to ≤ 1.6: moderately increased; > 1.6: severely increased. Female or unknown sex: ≤ 0.9: normal; > 0.9 to ≤ 1.2: mildly increased; > 1.2 to ≤ 1.5: moderately increased; > 1.5: severely increased.
CMR ventricular volumes, stroke volume, and mass	
LV end diastolic volume	If body surface area is available, use indexed range: male 50–108: normal; female 50–96: normal. Otherwise use absolute range: male 95–215: normal; female 78–167: normal. Below range: small; above range: dilated.
LV end systolic volume	If body surface area is available, use indexed range: male 11–47: normal; female 10–40: normal. Otherwise use absolute range: male 23–100: normal; female 18–73: normal. Below range: small; above range: dilated.
LV stroke volume	If body surface area is available, use indexed range: male 33–72: normal; female 33–64: normal. Otherwise use absolute range: male 60–135: normal; female 47–99: normal. Below range: reduced; above range: increased.
LV myocardial mass	If body surface area is available, use indexed range: male 39–85: normal; female 30–68: normal. Otherwise use absolute range: male 73–171: normal; female 61–121: normal. Below range: decreased; above range: increased.
RV end diastolic volume	If body surface area is available, use indexed range: male 53–123: normal; female 48–104: normal. Otherwise use absolute range: male 105–258: normal; female 84–193: normal. Below range: small; above range: dilated.
RV end systolic volume	If body surface area is available, use indexed range: male 17–59: normal; female 13–48: normal. Otherwise use absolute range: male 34–123: normal; female 22–86: normal. Below range: small; above range: dilated.
RV stroke volume	If body surface area is available, use indexed range: male 28–75: normal; female 29–66: normal. Otherwise use absolute range: male 57–141: normal; female 44–116: normal. Below range: reduced; above range: increased.
CMR atrial volumes and function	

Measurement	Categorisation rule
LA maximum volume	If body surface area is available, use indexed range: male 17–59: normal; female 17–61: normal. Otherwise use absolute range: male 31–112: normal; female 28–100: normal. Below range: small; above range: dilated.
LA minimum volume	If body surface area is available, use indexed range: male 3–24: normal; female 4–23: normal. Otherwise use absolute range: male 6–44: normal; female 7–38: normal. Below range: small; above range: dilated.
LA stroke volume	If body surface area is available, use indexed range: male 10–34: normal; female 10–34: normal. Otherwise use absolute range: male 21–67: normal; female 21–62: normal. Below range: reduced; above range: increased.
LA ejection fraction	Male 43–75: normal; female 47–75: normal. Below range: reduced; above range: hyperdynamic.
RA maximum volume	If body surface area is available, use indexed range: male 32–79: normal; female 31–69: normal. Otherwise use absolute range: male 59–158: normal; female 49–122: normal. Below range: small; above range: dilated.
RA minimum volume	If body surface area is available, use indexed range: male 10–42: normal; female 8–32: normal. Otherwise use absolute range: male 16–84: normal; female 11–55: normal. Below range: small; above range: dilated.
RA stroke volume	If body surface area is available, use indexed range: male 14–46: normal; female 14–42: normal. Otherwise use absolute range: male 26–90: normal; female 23–71: normal. Below range: reduced; above range: increased.
RA ejection fraction	Male 34–74: normal; female 41–77: normal. Below range: reduced; above range: hyperdynamic.
Global strain	
LV circumferential strain global	Male −27.2 to −14.6: normal; female −29.2 to −16.2: normal. Value above the upper limit: reduced magnitude; value below the lower limit: increased magnitude.
LV longitudinal strain global	Male −26.1 to −12.7: normal; female −28.7 to −14.2: normal. Value above the upper limit: reduced magnitude; value below the lower limit: increased magnitude.
LV radial strain global	Male and female 20.0–55.0: normal. Below range: reduced magnitude; above range: increased magnitude.
Aortic distensibility	
Ascending aorta distensibility	Male age $\leq$ 54: 0.7–5.1: normal; male age 55–64: 0.0–4.2: normal; male age $>$ 64: 0.0–2.4: normal. Female age $\leq$ 54: 0.5–5.7: normal; female age 55–64: 0.0–3.9: normal; female age $>$ 64: 0.0–2.7: normal. Below range: decreased; above range: increased.
Descending aorta distensibility	Male age $\leq$ 54: 1.6–6.0: normal; male age 55–64: 0.7–5.1: normal; male age $>$ 64: 0.4–4.0: normal. Female age $\leq$ 54: 1.6–6.0: normal; female age 55–64: 0.6–4.6: normal; female age $>$ 64: 0.4–3.6: normal. Below range: decreased; above range: increased.
EchoNext binary echocardiography rules	
LVEF $\leq$ 45%	Binary descriptor: yes if left ventricular ejection fraction is 45% or lower; otherwise no.
LV wall thickness $\geq$ 1.3 cm	Binary descriptor: yes if left ventricular wall thickness is at least 1.3 cm; otherwise no.
Aortic stenosis $\geq$ moderate	Binary descriptor: yes if aortic stenosis is moderate or greater; otherwise no.
Aortic regurgitation $\geq$ moderate	Binary descriptor: yes if aortic regurgitation is moderate or greater; otherwise no.
Mitral regurgitation $\geq$ moderate	Binary descriptor: yes if mitral regurgitation is moderate or greater; otherwise no.
Tricuspid regurgitation $\geq$ moderate	Binary descriptor: yes if tricuspid regurgitation is moderate or greater; otherwise no.
Pulmonary regurgitation $\geq$ moderate	Binary descriptor: yes if pulmonary regurgitation is moderate or greater; otherwise no.
RV systolic dysfunction $\geq$ moderate	Binary descriptor: yes if right ventricular systolic dysfunction is moderate or greater; otherwise no.
Pericardial effusion moderate or large	Binary descriptor: yes if pericardial effusion is moderate or large; otherwise no.
PASP $\geq$ 45 mmHg	Binary descriptor: yes if pulmonary artery systolic pressure is at least 45 mmHg; otherwise no.
TR maximum velocity $\geq$ 3.2 m/s	Binary descriptor: yes if tricuspid regurgitation maximum velocity is at least 3.2 m/s; otherwise no.

11 ECG-LLM Chat Template

During ECG-LLM instruction tuning and evaluation, each example was serialized as a chat conversation with a fixed system message, a user message containing the ECG token placeholder and clinical question, and an assistant message containing the target answer. The ECG token placeholder denotes the projected ECG-token sequence inserted into the LLM input embedding stream.

ECG-LLM chat template

```
System:
You are a cardiology assistant. The user will provide ECG data after the phrase "Here is the
↪  ECG:" followed by a clinical question. Answer the question directly using only findings
↪  supported by the ECG and available patient context. Do not invent unsupported findings,
↪  reason only on the provided ECG.

User:
Here is the ECG: <ecg_tokens>
<question_tokens>

Assistant:
<answer_tokens>
```

12 Question Answer Generation Prompts

The following prompts were used to generate question–answer pairs for training. Dataset-specific prompt variants were used for UK Biobank, EchoNext, MIMIC-IV-ECG, and PTB-XL according to the structured fields available in each cohort. For UK Biobank and EchoNext, we additionally generated reasoning-style QA pairs. In these examples, the generator returned separate `reasoning` and `answer` fields. During training, the model input was the generated question, and the target response was formed by concatenating the reasoning field with the final answer field.

Prompt for standard QA generation with UK Biobank data

```
Create N question-answer pairs from this text for LLM training in English.

BEHAVIOR:
- Simulate an interaction via QA pairs between a General Practitioner (GP) and a perfect AI
↪  system for ECG interpretation.
- The GP asks questions about this ECG, and the AI answers.
- The GP sees only the ECG and asks clinically meaningful, open-ended questions about this
↪  patient to obtain a comprehensive understanding of the patient's cardiac condition and
↪  abnormalities.
- The AI may internally use all provided information from the text.
- The AI must always answer as if the insights come purely from ECG interpretation.
- Answers should use as much provided information as possible, combining several fields to
↪  provide a comprehensive picture of the patient's cardiac status and abnormalities.
- Do not invent diagnoses, numbers, thresholds, or advice not supported by the provided
↪  fields.
- These QA pairs are intended for LLM fine-tuning. Across the full set of generated pairs,
↪  they should collectively reflect the patient's overall cardiac state, major
↪  abnormalities, and all clinically relevant supported findings in the text.

QUESTIONS (GP):
- Questions must sound like a GP asking about the ECG of this specific patient.
- Questions must not state or assume specific ECG diagnoses, measurements, or values.
- Do not ask about information that is not present in the text.
- Do not include the findings in the question. For example, do not ask: "If sinus bradycardia
↪  is present...", "Given LV dilatation...", or "If EF is reduced...".
- Instead, ask the AI to interpret, describe, and highlight abnormalities and to identify all
↪  relevant findings.
- Questions may reflect uncertainty or clinical curiosity, for example: "Does the ECG suggest
↪  ...?", "Could this ECG indicate ...?", "Do you see any signs of ...?", or "Is there
↪  evidence of ...?".

ANSWERS (AI):
- The AI has access to all information in the text, but must formulate answers as if the
↪  insights come purely from ECG interpretation.
- Answers must contain several clear sentences, with a direct and unambiguous answer to the
↪  question.
- Answers must include as much relevant information as possible about the patient's cardiac
↪  condition and abnormalities, combining several fields where appropriate.
- All answers must be directly supported by the provided fields and must never contradict any
↪  given measurements.
- Never mention anything not present in the provided text.

OUTPUT FORMAT (STRICT):
Return only valid JSON in this exact format:

[
  {{
    "question": "Question 1?",
    "answer": "Answer 1."
  }},
```

```
  {{
    "question": "Question 2?",
    "answer": "Answer 2."
  }}
]

Text:
{text}
```

Prompt for reasoning-style QA generation with UK Biobank data

```
Create N complex reasoning-style question-answer examples from this text for LLM training in
↪  English.

BEHAVIOR:
- Simulate an interaction via QA pairs between a General Practitioner (GP) and a perfect AI
↪  system for ECG interpretation.
- The GP asks questions about this ECG, and the AI answers.
- The GP sees only the ECG and asks clinically meaningful, open-ended questions about this
↪  patient to obtain a comprehensive understanding of the patient's cardiac condition and
↪  abnormalities.
- The AI may internally use all provided information from the text.
- The AI must always answer as if the insights come purely from ECG interpretation.
- Answers should use as much provided information as possible, combining several fields to
↪  provide a comprehensive picture of the patient's cardiac status and abnormalities.
- Do not invent diagnoses, numbers, thresholds, or advice not supported by the provided
↪  fields.
- These examples are intended for LLM fine-tuning. Across the full set of generated examples,
↪  they should collectively reflect the patient's overall cardiac state, major
↪  abnormalities, and all clinically relevant supported findings in the text.

QUESTIONS (GP):
- Questions must sound like a GP asking about the ECG of this specific patient.
- Questions must be clinically meaningful and require multi-step interpretation.
- Questions must not state or assume specific ECG diagnoses, measurements, or values.
- Do not ask about information that is not present in the text.
- Do not include the findings in the question. For example, do not ask: "If sinus bradycardia
↪  is present...", "Given LV dilatation...", or "If EF is reduced...".
- Instead, ask the AI to interpret, describe, and highlight abnormalities and to identify all
↪  relevant findings.
- Questions may reflect uncertainty or clinical curiosity, for example: "Does the ECG suggest
↪  ...?", "Could this ECG indicate ...?", "Do you see any signs of ...?", or "Is there
↪  evidence of ...?".

ANSWERS (AI):
- The AI has access to all information in the text, but must formulate answers as if the
↪  insights come purely from ECG interpretation.
- Each example must contain:
  1. a challenging ECG-focused clinical question,
  2. a concise reasoning field that links the available findings step by step,
  3. a final answer that directly addresses the question.
- The reasoning must be clinically grounded and directly supported by the provided fields.
- The final answer must be concise, direct, and unambiguous.
- All reasoning and answers must be directly supported by the provided fields and must never
↪  contradict any given measurements.
- Never mention anything not present in the provided text.

OUTPUT FORMAT (STRICT):
Return only valid JSON in this exact format:

[
  {{
    "question": "Complex question about this ECG?",
```

```
    "reasoning": "Step 1: ... Step 2: ... Step 3: ...",
    "answer": "Final answer based on the reasoning."
  }},
  {{
    "question": "Another complex question about this ECG?",
    "reasoning": "Step 1: ... Step 2: ... Step 3: ...",
    "answer": "Final answer based on the reasoning."
  }}
]

Text:
{text}
```

Prompt for QA generation with MIMIC-IV-ECG data

```
Create N question-answer pairs from this text for LLM training in English.

BEHAVIOR:
- Simulate an interaction between a General Practitioner (GP) and a PERFECT AI for ECG
↪  interpretation.
- The GP asks questions about this ECG, and the PERFECT AI answers.
- The GP sees only the ECG and asks clinically meaningful, open-ended questions about THIS
↪  patient to obtain a comprehensive understanding of the patient's heart condition and
↪  abnormalities.
- PERFECT AI for ECG interpretation may internally use and reason with ALL provided
↪  information from the text (ECG phenotypes: + Patient demographics),
  but MUST answer as if insights come purely from perfect ECG interpretation.
- Questions and answers must be phrased as if inferred from ECG alone. Never mention records,
↪  reports, metadata, or models.
- Do not invent diagnoses, numbers, thresholds, or advice not supported by the provided
↪  fields.
- These QA pairs are intended for LLM fine-tuning. Across the full set of generated pairs,
↪  they should collectively reflect the patient's overall cardiac state, major
↪  abnormalities, and all clinically relevant supported findings in the text.

QUESTIONS (GP):
- Must sound like a GP interpreting and asking questions about the ECG of this specific
↪  patient.
- MUST NOT state or assume specific ECG diagnoses, measurements, or values in the question.
- Do not ask about information that is not present in the text.
- Do not include the findings in the question. Instead, ask the AI to interpret, describe,
↪  and highlight abnormalities and to explain all relevant findings.
- May reflect uncertainty or clinical curiosity: "Does the ECG suggest ...?", "Could this ECG
  indicate...?", "Do you see any signs of...", "Is there evidence of...", etc.

ANSWERS (AI):
- This is a PERFECT AI for ECG interpretation. It has access to all information in the text,
↪  but must formulate answers as if insights come purely from ECG interpretation.
- Answers must be several clear sentences, with a direct and unambiguous answer to the
↪  question.
- Answers must include as much relevant information as possible about the patient's heart
↪  condition and abnormalities, combining several fields.
- All answers must be directly supported by the provided fields and must never contradict any
↪  given measurements.
- Never mention anything not present in the provided text.

OUTPUT FORMAT (STRICT):
Return ONLY valid JSON in this exact format:

[
  {{
    "question": "Question 1?",
    "answer": "Answer 1."
```

```
  }},
  {{
    "question": "Question 2?",
    "answer": "Answer 2."
  }}
]

Text:
{text}
```

Prompt for QA generation with PTB-XL data

```
Create N question-answer pairs from this text for LLM training in English.

BEHAVIOR:
- Simulate an interaction between a General Practitioner (GP) and a PERFECT AI for ECG
↪  interpretation.
- The GP asks questions about this ECG, and the PERFECT AI answers.
- The GP sees only the ECG and asks clinically meaningful, open-ended questions about THIS
↪  patient to obtain a comprehensive understanding of the patient's heart condition and
↪  abnormalities.
- PERFECT AI for ECG interpretation may internally use and reason with ALL provided
↪  information from the text.
- PERFECT AI for ECG interpretation ALWAYS answer as if insights come purely from ECG
↪  analysis.
- The output QA pairs MUST be ONLY in English.
- Do not invent diagnoses, numbers, thresholds, or advice not supported by the provided
↪  fields.
- These QA pairs are intended for LLM fine-tuning. Across the full set of generated pairs,
↪  they should collectively reflect the patient's overall cardiac state, major
↪  abnormalities, and all clinically relevant supported findings in the text.
- QA pairs must cover all findings from the text.

QUESTIONS (GP):
- Must sound like a GP interpreting and asking questions about the ECG of this specific
↪  patient.
- NEVER include the findings from the text in the question. For example, do NOT say "If sinus
↪  bradycardia is present...", "Given LV dilatation...", or "If EF is reduced...". Instead,
↪  ask the AI to interpret, describe, and highlight abnormalities and to explain all
↪  relevant findings.
- May reflect uncertainty or clinical curiosity: "Does the ECG suggest ...?", "Could this ECG
  indicate...?", "Do you see any signs of...", "Is there evidence of...", etc.
- Do not ask about information that is not present in the text.

ANSWERS (AI):
- This is a PERFECT AI for ECG interpretation. It must formulate answers as if insights come
↪  purely from ECG interpretation.
- Answers must be several clear sentences, with a direct and unambiguous answer to the
↪  question.
- Answers must include as much relevant information as possible about the patient's heart
↪  condition and abnormalities, combining several fields.
- All answers must be directly supported by the provided fields and must never contradict any
↪  given measurements.
- Never mention anything not present in the provided text.

OUTPUT FORMAT (STRICT):
Return ONLY valid JSON in this exact format:

[
  {{
    "question": "Question 1?",
    "answer": "Answer 1."
  }},
```

```
  {{
    "question": "Question 2?",
    "answer": "Answer 2."
  }}
]

Text:
{text}
```

Prompt for QA generation with EchoNext data

```
Create N question-answer pairs from this text for LLM training in English.

BEHAVIOR:
- Simulate an interaction via QA pairs between a General Practitioner (GP) and a PERFECT AI
↪  for ECG interpretation.
- The GP asks questions about this ECG, and the PERFECT AI answers.
- The GP sees only the ECG and asks clinically meaningful, open-ended questions about THIS
↪  patient to obtain a comprehensive understanding of the patient's heart condition and
↪  abnormalities.
- PERFECT AI for ECG interpretation may internally use and reason with ALL provided
↪  information from the text.
- PERFECT AI for ECG interpretation ALWAYS answer as if insights come purely from ECG.
- Answers should cite as much provided information as possible, combing several fields to
↪  provide a comprehensive picture of the patient's cardiac status and abnormalities.
- Do not invent diagnoses, numbers, thresholds, or advice not supported by the provided
↪  fields.
- These QA pairs are intended for LLM fine-tuning. Across the full set of generated pairs,
↪  they should collectively reflect the patient's overall cardiac state, major
↪  abnormalities, and all clinically relevant supported findings in the text.

QUESTIONS (GP):
- Must sound like a GP asking questions about the ECG of this specific patient. GP asks
↪  questions.
- MUST NOT state or assume specific ECG diagnoses, measurements, or values in the question.
- Do not ask about information that is not present in the text.
- Do not include the findings in the question. For example, do NOT say "If sinus bradycardia
↪  is present...", "Given LV dilatation...", or "If EF is reduced...".
- Instead, ask the AI to interpret, describe, and highlight abnormalities and to find all
↪  relevant abnormalities.
- Questions may reflect uncertainty or clinical curiosity: "Does the ECG suggest ...?",
↪  "Could this ECG
  indicate...?", "Do you see any signs of...", "Is there evidence of...", etc.

ANSWERS (AI):
- This is a PERFECT AI for ECG interpretation. It has access to all information in the text,
↪  and it must formulate answers as if insights come purely from ECG interpretation.
- Answers must be several clear sentences, with a direct and unambiguous answer to the
↪  question.
- Answers must include as much relevant information as possible about the patient's heart
↪  condition and abnormalities, combining several fields.
- All answers must be directly supported by the provided fields and must never contradict any
↪  given measurements.
- Never mention anything not present in the provided text.

OUTPUT FORMAT (STRICT):
Return ONLY valid JSON in this exact format:

[
  {{
    "question": "Question 1?",
    "answer": "Answer 1."
  }},
```

```
  {{
    "question": "Question 2?",
    "answer": "Answer 2."
  }}
]

Text:
{text}
```

Prompt for reasoning-style QA generation with EchoNext data

```
Create N complex reasoning examples from this text that demonstrate chain-of-thought
↪  thinking.

BEHAVIOR:
- Simulate an interaction via QA pairs between a General Practitioner (GP) and a PERFECT AI
↪  for ECG interpretation.
- The GP asks questions about this ECG, and the PERFECT AI answers.
- The GP sees only the ECG and asks clinically meaningful, open-ended questions about THIS
↪  patient to obtain a comprehensive understanding of the patient's heart condition and
↪  abnormalities.
- PERFECT AI for ECG interpretation may internally use and reason with ALL provided
↪  information from the text.
- PERFECT AI for ECG interpretation ALWAYS answer as if insights come purely from ECG.
- Answers should cite as much provided information as possible, combing several fields to
↪  provide a comprehensive picture of the patient's cardiac status and abnormalities.
- Do not invent diagnoses, numbers, thresholds, or advice not supported by the provided
↪  fields.
- These QA pairs are intended for LLM fine-tuning. Across the full set of generated pairs,
↪  they should collectively reflect the patient's overall cardiac state, major
↪  abnormalities, and all clinically relevant supported findings in the text.

QUESTIONS (GP):
- Must sound like a GP asking questions about the ECG of this specific patient. GP asks
↪  questions.
- MUST NOT state or assume specific ECG diagnoses, measurements, or values in the question.
- Do not ask about information that is not present in the text.
- Do not include the findings in the question. For example, do NOT say "If sinus bradycardia
↪  is present...", "Given LV dilatation...", or "If EF is reduced...".
- Instead, ask the AI to interpret, describe, and highlight abnormalities and to find all
↪  relevant abnormalities.
- Questions may reflect uncertainty or clinical curiosity: "Does the ECG suggest ...?",
↪  "Could this ECG
  indicate...?", "Do you see any signs of...", "Is there evidence of...", etc.

ANSWERS (AI):
- This is a PERFECT AI for ECG interpretation. It has access to all information in the text,
↪  and it must formulate answers as if insights come purely from ECG interpretation.
- Answers must be several clear sentences, with a direct and unambiguous answer to the
↪  question.
- Answers must include as much relevant information as possible about the patient's heart
↪  condition and abnormalities, combining several fields.
- All answers must be directly supported by the provided fields and must never contradict any
↪  given measurements.
- Never mention anything not present in the provided text.

Each example should have:
1. A challenging question that requires step-by-step reasoning
2. Detailed reasoning steps that break down the problem
3. A concise final answer

Return JSON format only:
```

```
[
  {{
    "question": "Complex question about the text?",
    "reasoning": "Step 1: First, I need to consider...\nStep 2: Then, I analyze...\nStep 3:
    ↪  Finally, I can conclude...",
    "answer": "Final answer based on the reasoning."
  }},
  {{
    "question": "Another complex question?",
    "reasoning": "First, I'll analyze... Next, I need to determine... Based on this
    ↪  analysis...",
    "answer": "Final answer drawn from the reasoning."
  }}
]

Text:
{text}
```

13 Implementation Details

Parameter	Value
Input	12 leads × 5,000 samples
Patch size	1×100
ECG tokens	300 before masking
Masking ratio	0.75
Loss	MSE over masked patches
Encoder	ViT, 12 layers, hidden size 768, 12 heads
Decoder	Transformer, 2 layers, hidden size 512, 8 heads
Activation	GELU
Positional embedding	Learnable
Drop-path rate	0.1
Optimiser	AdamW
Learning rate	2×10^{-4}
Weight decay	0.05
Batch size	768
Precision	bfloat16 mixed precision
Training duration	120 epochs
Warm-up	1,000 steps
Minimum LR factor	0.01
Augmentations	Random crop, Fourier surrogate perturbation, jitter, rescaling

Table 11: ECG ViT-MAE pretraining configuration.

Parameter	Value
ECG encoder	Pretrained ViT-MAE, frozen
Language model	Llama-3.1-8B-Instruct, frozen backbone
ECG crop	Random crop to 2,500 time-steps
ECG augmentations	Disabled
Projector	Linear($d_{\text{ecg}}, d_{\text{llm}}$), SiLU, Linear($d_{\text{llm}}, d_{\text{llm}}$)
Aggregation	Identity; all ECG tokens preserved
Projector bias	Disabled
LoRA modules	`q_proj`, `k_proj`, `v_proj`, `o_proj`
LoRA rank / alpha	128 / 128
LoRA dropout	0.1
Curriculum	First 1,500 steps update only the projector, then projector + LoRA training
Optimiser	Paged AdamW 8-bit
Learning rates	Projector 10^{-4}; LoRA 10^{-5}
Weight decay	10^{-3}
Schedule	Linear warm-up from 0.1× to peak LR over 3% of optimiser steps, then cosine decay to 0.3× peak LR
Precision	bfloat16 mixed precision
Training duration	1 epoch
Batching	4 GPUs, batch size 8/GPU, gradient accumulation 4
Effective batch size	128
Max sequence length	1200 tokens

Table 12: ECG-LLM instruction-tuning configuration.

14 Diagnostic report generation prompts

Model	Diagnostic report generation prompt
ECG-LLM	Review the ECG signal and produce a detailed report on your diagnostic observations, ending with the final diagnosis. Describe this ECG in detail and tell if there are any abnormalities present. Review this ECG and provide an ECG diagnosis. What is wrong with this ECG? Provide an ECG report for this tracing. What are the key diagnostic findings in this ECG? Analyze this ECG carefully and provide a detailed ECG report. What is the summary for this ECG?
ECG-Chat	Could you please help me explain my ECG?
PULSE-7B	Please write a clinical report based on this ECG image.
MedGemma-4B	Please interpret this 12-lead ECG image. Provide a concise diagnostic report, including rhythm, rate if assessable, conduction or axis abnormalities if present, ST-T or infarction-related findings if present, and an overall ECG impression. If a finding is uncertain or not clearly visible, say so rather than inventing it.

Table 13: Prompts used for ECG diagnostic summary generation. For ECG-LLM, one question was randomly selected from the ECG-LLM prompt pool for each ECG. For baseline models, the model-specific prompt shown in the table was used.

15 LLM-as-judge prompt for ECG diagnosis evaluation

The following prompts were used for LLM-based evaluation of generated ECG diagnostic answers against ground-truth diagnostic statements.

System prompt for LLM-as-judge evaluation

```
You are a strict but clinically fair ECG diagnosis evaluator.

Compare MODEL ANSWER against GROUND TRUTH. Treat clinically equivalent wording as correct. Do not
↪  require identical phrasing.
Score correctness, completeness, and specificity independently.

Clinical equivalence rules:
- Borderline ECG, abnormal ECG, nonspecific, possible, probable, normal variant, and age-undetermined
↪  findings are lower-confidence/global findings.
- Do not strongly penalize missing uncertain findings if the main ECG interpretation is correct.
- Penalize missing GT findings under completeness and unsupported extra findings under specificity. Do
↪  not assign 0 if at least one meaningful GT finding is correctly identified. But don't assign 2 if
↪  important GT findings are missing or if the model adds unsupported major findings.

For PTB-XL ground truth: #Added only for PTB-XL evaluation
- Use the Original report as the primary reference.
- Use the English report only as a translation aid.
- Do not treat the English report as an additional independent list of required findings.

Return only valid JSON.
```

User prompt template for LLM-as-judge evaluation

```
GROUND TRUTH:
{gt}

MODEL ANSWER:
{pred}

Score independently.

Correctness:
2 = main ECG interpretation is clinically correct or nearly correct.
1 = partially correct: some main findings are correct, but there is an important wrong or missing
↪  conclusion.
0 = mostly incorrect, or the main diagnosis conclusion is wrong.

Completeness:
2 = all or nearly all important GT diagnoses are present.
1 = at least one important GT diagnosis is present, but relevant GT findings are missing.
0 = no clinically meaningful GT diagnosis is present.

Specificity:
2 = nearly all model diagnoses are supported by or compatible with GT.
1 = mixed answer: some supported diagnoses plus unsupported or wrong extras.
0 = mostly unsupported or dominated by wrong diagnoses.

Return JSON exactly:
{{
  "correctness": int,
  "completeness": int,
  "specificity": int,
  "total_score": int,
  "notes": "reason"
}}
```

References

[1] Akshay Goel. Langextract v1.4.0, 2026. Version 1.4.0.

[2] Google. Langextract. `https://github.com/google/langextract`, 2025. GitHub repository.

[3] OpenAI. Introducing gpt-oss. https://openai.com/index/introducing-gpt-oss/, 2025.

[4] OpenAI. gpt-oss-20b model card. https://huggingface.co/openai/gpt-oss-20b, 2025.

[5] OpenAI. gpt-oss-120b and gpt-oss-20b model card, 2025.