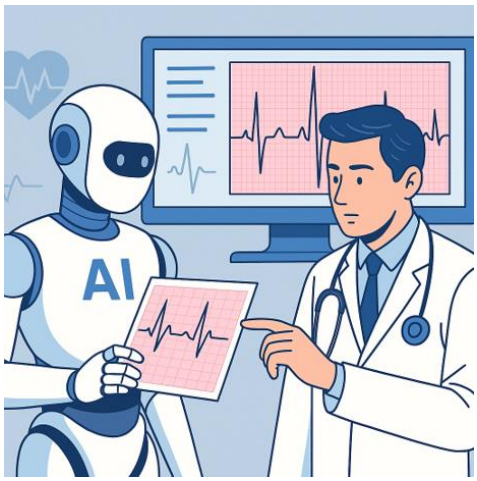
NATIONAL TECHNICAL UNIVERSITY OF ATHENS

SCHOOL OF ELECTRICAL AND COMPUTER ENGINEERING

SCHOOL OF MECHANICAL ENGINEERING

INTERDISCIPLINARY POSTGRADUATE PROGRAMME

“Translational Engineering in Health and Medicine”



Optimizing Cardiology Diagnostic Accuracy via Synergistic AI-Human Integration

Postgraduate Diploma Thesis

Michael Kalogeropoulos

Supervisor: ***Konstantina (Nantia) S. Nikita,*** *M.Eng., M.D., Ph.D.*
Professor
Editor-in-Chief, IEEE Transactions on Antennas and Propagation
School of Electrical and Computer Engineering
National Technical University of Athens

Athens, July 2025

NATIONAL TECHNICAL UNIVERSITY OF ATHENS

SCHOOL OF ELECTRICAL AND COMPUTER ENGINEERING

SCHOOL OF MECHANICAL ENGINEERING

INTERDISCIPLINARY POSTGRADUATE PROGRAMME

"Translational Engineering in Health and Medicine"

Optimizing Cardiology Diagnostic Accuracy via Synergistic AI-Human Integration

Postgraduate Diploma Thesis

Michael Kalogeropoulos

Supervisor: ***Konstantina (Nantia) S. Nikita,*** *M.Eng., M.D., Ph.D.*
Professor
Editor-in-Chief, IEEE Transactions on Antennas and Propagation
School of Electrical and Computer Engineering
National Technical University of Athens

The postgraduate diploma thesis has been approved by the examination committee on 04/12/2024 *(exam day: 04 July 2025)*

1st member	*2nd member*	*3rd member*
Prof. Konstantina S. Nikita	*Prof. Giorgos Stamou*	*Assist. Prof. Athanasios (Thanos) Voulodimos*
NTUA / School of Electrical and Computer Engineering	*NTUA / School of Electrical and Computer Engineering*	*NTUA / School of Electrical and Computer Engineering*

Athens, July 2025

Michael Kalogeropoulos
Graduate of the Interdisciplinary Postgraduate Programme,
“Translational Engineering in Health and Medicine”,
Master of Science,
School of Electrical and Computer Engineering,
National Technical University of Athens

Abstract

Introduction: Artificial intelligence (AI) is becoming an increasingly important part of healthcare, especially as a tool to support clinical decision-making. Electrocardiogram (ECG) is a routine but critical task in medical practice, making it a strong candidate for AI assistance. While AI models like GPT-4o have shown strong accuracy in diagnosing ECGs, there's still little real-world evidence on how they actually influence doctors' decisions during interpretation. This study aims to close that gap by examining how GPT-4o affects clinicians' diagnostic accuracy when reading ECGs.

Methods: We carried out a controlled study using a questionnaire-based design with 25 physicians at different levels of experience: 10 cardiologists, 10 experienced internists, and 5 less experienced internists. Each participant reviewed the same set of 50 ECG cases twice, first on their own and then with GPT-4o's diagnostic suggestion for each case. The cases included 20 everyday ECGs, 20 more challenging ones, and 10 extra, challenging ECGs cases where the AI provided an intentionally incorrect suggestion to test whether it could lead physicians into making errors. For each case, physicians recorded their initial diagnosis and had the option revising it after seeing the AI's suggestion. We compared diagnostic accuracy with and without AI assistance across experience levels and case difficulty, and we also tracked how often participants changed their answers. Statistical tests were used to validate whether these differences were significant.

Results: GPT-4o achieved an accuracy of 72.5% on the ECG cases. With AI assistance, diagnostic accuracy improved for all physician groups. Cardiologists improved from 81.6% without AI to 84.8% with it. Experienced internists saw their accuracy rise from 63.4% to 73.8%, while less experienced internists improved from 43.2% to 55.6%. The biggest improvement was seen in the most difficult cases, where less experienced internists jumped from 38% to 75% with AI support. Overall, physicians mostly used AI to fix their initial mistakes, wrong-to-right answer changes outnumbered right-to-wrong changes by about 4:1. However, in cases with deliberately misleading AI suggestions, less experienced internists were completely misled, with their accuracy dropping to 0% in those cases due to overreliance on the AI. However, in cases where the AI gave intentionally incorrect suggestions, less experienced internists were fully misled, dropping to 0% accuracy due to overreliance. All improvements in accuracy were statistically significant ($p < 0.05$), although the size of the benefit varied depending on experience level.

Conclusion: GPT-4o significantly improved ECG interpretation accuracy across all physician groups, with the biggest gains seen among less experienced clinicians. Even senior doctors saw modest benefits. However, the study also highlights the risk of automation bias when the AI was wrong, less experienced physicians were especially prone to follow its suggestions, leading to major drops in accuracy. These findings show that while AI like GPT-4o can be a valuable diagnostic aid, it must be used with caution.

Keywords: Artificial Intelligence, Large Language Models, OpenAI, GPT-4o, Electrocardiography, Diagnostic Accuracy, Human-AI Interaction, Clinical Decision-Making, AI co-pilot

Acknowledgements

I would like to express my sincere gratitude to Professor Konstantina S. Nikita for giving me the opportunity to join the Biomedical Simulations and Imaging Laboratory at the School of Electrical and Computer Engineering, National Technical University of Athens. Her guidance and vision have been instrumental in making this thesis possible.

I am deeply thankful to Dr. Konstantinos Mitsis for his continuous support, encouragement, and close supervision throughout this work, as well as for his invaluable contributions to the design and implementation of the associated medical research.

Special thanks go to my classmate Katerina, for her collaboration, support and teamwork across numerous projects and courses throughout our postgraduate journey.

Finally, I would like to thank my family, my parents Theano and Dimitrios, my brother Stamatios, and my uncles Giannis and Petros for their unwavering support, love, encouragement and belief in me every step of the way.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

GLOSSARY

AI co-pilot: Refers to the concept of an AI system acting as a "co-pilot" alongside a human expert in decision-making. In this thesis, it describes AI (like GPT-4o) assisting clinicians in interpreting ECGs, providing suggestions and analysis while the human remains the final decision-maker. An AI co-pilot is intended to handle routine tasks or offer a second opinion, augmenting human capabilities rather than replacing them.

Arrhythmia: Any abnormal heart rhythm. This broad term covers irregularities in the heartbeat's rate or pattern, ranging from benign extra beats to serious conditions like fibrillation or tachycardia. Arrhythmias can affect blood flow and are often diagnosed via ECG tracings by identifying irregular wave patterns.

Artificial intelligence: A field of computer science focused on creating machines or software that can perform tasks requiring human-like intelligence. These tasks include learning from data, recognizing patterns, making decisions, and predicting outcomes. In healthcare, artificial intelligence (often abbreviated AI) is used for tasks like interpreting medical images or signals and supporting clinical decisions through pattern recognition and data analysis.

Atrial fibrillation: A common cardiac arrhythmia characterized by rapid, irregular beating of the atria (the heart's upper chambers). It results in an erratic pulse and inefficient blood flow, increasing the risk of stroke and other complications. On an ECG, atrial fibrillation (abbreviated AF) shows an absence of distinct P waves and an irregularly irregular QRS rhythm.

Augmented intelligence: Also known as the "team advantage," this concept refers to the enhanced performance achieved when humans and AI work together compared to either alone. Rather than AI replacing human intelligence, augmented intelligence emphasizes AI as a tool that complements and extends human decision-making. For example, a doctor and an AI assistant together may catch diagnostic details or errors that one might miss individually, leading to improved accuracy.

Automation bias: The tendency to trust and rely too heavily on automated systems or AI suggestions, even when they may be wrong. In a clinical setting, automation bias can cause a clinician to accept an AI's diagnosis or advice without sufficient scrutiny, potentially overriding their own correct judgment. This bias is especially risky if the AI output is incorrect, as it can lead to diagnostic errors due to overreliance on the technology.

Bias (AI): In the context of AI, bias refers to systematic errors or unfairness in the model's outputs caused by prejudices or imbalances in the training data or algorithms. AI bias can manifest as favoring certain groups (e.g., along racial or gender lines) or consistently misrepresenting information. In healthcare, biased AI could lead to unequal or incorrect recommendations if not addressed through careful training and validation.

Black box (AI): A term describing AI systems (especially deep learning models) whose internal decision-making processes are not transparent or easily interpretable to humans. When an AI is a "black box," it produces outputs (e.g., a diagnosis or prediction) without providing insight into how it arrived at that result. This lack of explainability can hinder trust, as clinicians may be hesitant to accept recommendations if they cannot understand the rationale behind them.

Cardiologist: A physician specialized in cardiology, the branch of medicine dealing with the heart and circulatory system. Cardiologists have advanced training in diagnosing and treating heart conditions and are experts in interpreting ECGs and other cardiac tests. In this thesis, cardiologists (with 15–25 years of experience) represent the most experienced group in ECG interpretation, serving as a benchmark for expert-level performance.

Chi-square test: A statistical test used to compare observed frequencies to expected frequencies in categorical data, to see if there is a significant difference or association. In this thesis, chi-square tests (χ^2) were used to analyze categorical outcomes like the distribution of answer changes. A large chi-square statistic with a low p-value indicates that the differences in frequencies (for example, how often different groups changed their answers) are unlikely due to chance alone.

Confidence interval: In statistics, a confidence interval (CI) is a range of values, derived from sample data, that likely contains the true population value for a given measure with a certain level of confidence (often 95%). For example, a 95% confidence interval of [0.4, 0.8] for an improvement in accuracy means we can be 95% sure the true improvement lies between 40% and 80%. If a CI for a difference does not include zero, that difference is considered statistically significant (since zero would mean no difference).

Conformal prediction: A machine learning framework that produces predictions along with a measure of confidence in the form of prediction sets. Instead of giving a single answer, a conformal prediction model might output a set of possible diagnoses with associated confidence levels that the true answer lies within that set. This approach aims to provide calibrated uncertainty estimates for instance, listing a few likely diagnoses for an ECG along with probabilities to help users gauge how much to trust the AI's output and reduce overconfidence in any single suggestion.

Deep learning: A subset of machine learning that uses artificial neural networks with multiple layers (hence "deep") to learn complex patterns from large amounts of data. Deep learning has driven many advances in AI, particularly in image and signal interpretation. In the context of this thesis, deep learning models have been used to interpret ECG waveforms and can achieve expert-level accuracy in specific tasks (such as rhythm classification) by automatically learning features from the data.

Differential diagnosis: The process of compiling a list of potential conditions that could explain a patient's symptoms or findings, then narrowing it down to the most likely one. Clinicians use differential diagnoses to consider various possibilities before arriving at a final diagnosis. AI tools can assist by suggesting diagnoses that a clinician might include in their differential. For example, given an ECG and symptoms, the differential diagnosis might include several cardiac conditions, and an AI could help ensure no important possibility is overlooked.

Differential privacy: A technique in data science and AI designed to protect individual data points when aggregating or analyzing data. Differential privacy adds a controlled amount of noise to data or query results, so that the output does not reveal specifics about any one individual, thus preserving confidentiality. In the thesis context, approaches like differential privacy (along with federated or swarm learning) are discussed as ways to train AI models on sensitive medical data (like patient records or ECGs) without compromising patient privacy.

ECG telemetry: Continuous monitoring of a patient's heart electrical activity (ECG) in real time, often used in hospital settings (like an intensive care unit or telemetry ward). ECG telemetry allows for ongoing observation of heart rhythms to catch transient arrhythmias or ischemic changes that might occur between routine checks. An AI co-pilot integrated with ECG telemetry could analyze these continuous streams and alert clinicians to subtle or intermittent abnormalities that require attention.

ECGphobia: An informal term referring to the anxiety or lack of confidence that some medical trainees or non-cardiology doctors feel about interpreting electrocardiograms. ECG interpretation is complex, and "ECGphobia" captures the intimidation or fear of misreading ECGs. The thesis mentions this concept to highlight why decision support tools (like AI assistants) are appealing – they could help less experienced clinicians overcome their uncertainty by providing guidance or confirmation when reading ECGs.

Effect size: A quantitative measure of the magnitude of a result or difference, independent of sample size. Unlike a p-value, which only tells if an effect is statistically significant, an effect size indicates how large or meaningful that effect is in practical terms. Examples include Cohen's d for difference in means, ϕ (phi) for chi-square tests, or η^2 (eta-squared) for variance explained. In the study, large effect sizes (e.g., Cohen's d around 1.0 or higher) indicated that AI assistance had a very strong impact on improving accuracy or altering outcomes, beyond just being statistically significant.

Explainability (AI): The degree to which an AI system can provide understandable explanations for its decisions or predictions. High explainability means the model can reveal why it gave a certain output (for example, highlighting which ECG features led to a diagnosis or providing reasoning in natural language). Explainability is important for trust – clinicians are more likely to trust and effectively use an AI assistant if they can follow its line of reasoning. The thesis discusses the "black box" problem: many AI models (like deep neural networks) lack explainability, which can be a barrier to their adoption in clinical practice.

False confirmation: A scenario in human-AI interaction where both the human and the AI independently lean toward the same incorrect conclusion, thereby reinforcing each other's confidence in that wrong answer. For example, a physician might initially choose an incorrect diagnosis and then see an AI suggestion that echoes the same incorrect choice. This "confirmation" can falsely boost the clinician's confidence that the wrong answer is correct, since both human and AI agree. False confirmation is dangerous because it can cement a mistake, making the duo of human+AI overly confident in a flawed decision.

False conflict error: A cognitive error that occurs when a clinician initially has the correct answer but then changes to a wrong answer after receiving a conflicting suggestion from an AI. In this

situation, the AI's incorrect recommendation creates a "false conflict" with the clinician's correct judgment. The clinician, doubting their own answer, defers to the AI and ends up making an error they would not have made otherwise. The thesis found that less experienced doctors were highly susceptible to false conflict errors, often abandoning correct decisions when the AI disagreed.

Federated learning: A method of training AI models on data that is distributed across multiple devices or servers without centralizing the data. In federated learning, the model is sent to where the data reside (for instance, different hospitals or patient devices), learns from local data, and only the learned parameters are sent back and aggregated to update a global model. This way, sensitive data (like patient ECGs) never leaves its source. Federated learning can enhance privacy and security because raw data isn't pooled in one location, aligning with privacy needs in healthcare.

Fine-tuning: The process of taking a pre-trained machine learning model (often a large model trained on broad data) and training it further on a specific task or domain using a smaller, task-specific dataset. Fine-tuning adjusts the model's parameters to make it perform better on the targeted application. In the thesis context, a large language model like GPT-4 can be fine-tuned on medical data or ECG cases to improve its performance in clinical diagnosis. However, fine-tuning requires significant data and resources and, if done improperly, can introduce biases or overfitting.

Fleiss' kappa: A statistical measure of inter-rater reliability for agreement among three or more raters. It extends the concept of Cohen's kappa (which is for two raters) to multiple observers. Fleiss' kappa values range from 0 (no agreement beyond chance) to 1 (perfect agreement). In the thesis, a reference is made to a Fleiss' kappa of 0.51 for GPT-4o's consistency in another study, indicating moderate agreement among its multiple outputs. A moderate kappa suggests the AI's responses were reasonably consistent (important for a reliable assistant), although not perfectly identical each time.

G*Power: A statistical software tool used for power analysis and sample size calculation. Researchers use GPower to determine how many participants are needed to detect an expected effect size with a given level of statistical power. In this study, GPower (version 3.1.9.7) was used to perform an a priori power analysis it helped calculate that about 20 participants (10 per key group) would be sufficient to detect the anticipated differences in accuracy between physician groups with high power (≥95% chance of detecting the effect if it exists).

Gemini (Google's Gemini): The codename of a next-generation large AI model being developed by Google, mentioned as a counterpart or competitor to OpenAI's GPT-4 series. Gemini is expected to be a multimodal model (handling text, images, etc.) and to have powerful reasoning capabilities. In the thesis, Google's Gemini is noted in the context of future AI tools; early benchmarks suggested GPT-4o outperformed an initial version of Gemini on ECG tasks, but ongoing development means such comparisons will evolve. Essentially, "Gemini" represents the cutting-edge AI model from Google's research, analogous to how GPT-4 represents OpenAI's.

Human–AI collaboration: A working approach where human experts and AI systems interact and contribute jointly to a task, each complementing the other's strengths. In medical diagnosis,

human–AI collaboration could mean a doctor and an AI exchange information the AI offers suggestions or analysis (like pointing out an ECG abnormality), and the human validates and decides based on their clinical judgment and the AI input. The goal of this collaboration is to achieve better outcomes than either could alone (e.g., higher diagnostic accuracy, efficiency, or confidence). The thesis examines how such collaboration works in practice, including benefits (catching each other's mistakes) and pitfalls (overreliance or conflicting inputs).

Hypertrophic cardiomyopathy: A genetic heart condition characterized by abnormal thickening of heart muscle (usually the left ventricle). This thickening can lead to obstruction of blood flow and predispose to arrhythmia. Hypertrophic cardiomyopathy (HCM) is notable on an ECG for signs like very large voltage QRS complexes or deep inverted T-waves in certain leads, but it can be subtle. The thesis references this condition as an example of what AI algorithms have been trained to detect from ECGs, demonstrating that AI can find patterns of disease (like HCM) that might be difficult for humans to spot early.

Internist: Short for an internal medicine physician, an internist is a doctor specializing in the prevention, diagnosis, and treatment of adult diseases. Internists are not surgeons and usually are not organ-specific specialists (like cardiologists), but they manage a broad range of conditions. In this thesis, "experienced internists" refers to seasoned internal medicine doctors (15–25 years of experience) and "less experienced internists" to recent graduates (<5 years of experience). The performance and behavior of these internists in ECG interpretation (with and without AI help) are compared to those of cardiologists to understand how expertise level interacts with AI support.

Ischemia: A condition in which tissue (such as heart muscle) receives insufficient blood flow and oxygen, usually due to a blocked or narrowed artery. In the heart, ischemia often manifests as chest pain (angina) and specific changes on an ECG, like ST-segment depression or T-wave inversions. Prolonged severe ischemia can lead to myocardial infarction (heart muscle cell death). The thesis mentions ischemia in the context of ECG interpretation subtle signs of cardiac ischemia on an ECG can be life-saving clues (as prompt treatment can restore blood flow). Detecting ischemia is crucial, and both clinicians and AI tools aim to recognize its ECG patterns.

Kruskal–Wallis test: A non-parametric statistical test used to compare the distributions of a continuous or ordinal outcome across three or more independent groups. It is an alternative to one-way ANOVA when the data do not meet the assumptions of normality or equal variances. In the study, a Kruskal–Wallis test was used (for example, to compare overall accuracy among the three physician groups) when normality was violated for one of the groups. A significant Kruskal–Wallis result (reported with an H statistic and p-value) indicates that at least one group differs significantly from the others, prompting further post-hoc comparisons to identify which groups differ.

Large language model: A type of AI model that is trained on massive amounts of text data and has many parameters (often billions), enabling it to understand and generate human-like language. Large language models (LLMs) use advanced architectures (like the Transformer) and can perform a variety of language tasks from answering questions and summarizing text to reasoning and code generation. In healthcare, LLMs (e.g., GPT-4) can be applied to understand medical texts, patient records, or even describe medical images and signals. The thesis

specifically evaluates an LLM (GPT-4o) in the context of reading ECG cases and providing diagnostic suggestions, highlighting both its impressive performance and its challenges (like occasional "hallucinations" or lack of real-world experience).

Levene's test: A statistical test used to assess the equality of variances for a variable across different groups. Before performing an ANOVA, Levene's test helps determine whether the assumption of equal variance (homogeneity of variance) holds. A non-significant Levene's test ($p > 0.05$) means the variances are roughly equal and the assumption is met; a significant result means variances differ, and one might use a different approach or a more robust comparison (like Welch's ANOVA or non-parametric tests). In the thesis, Levene's test was applied when comparing physician groups' performances; for example, it was reported that variance assumptions were met (Levene's $p = 0.12$ in one case, indicating no significant difference in variances).

Machine learning: A branch of AI focused on algorithms that allow computers to learn patterns from data and improve performance on a task with experience. Instead of being explicitly programmed for every rule, machine learning models adjust their internal parameters based on training data. Techniques include supervised learning (learning from labeled examples), unsupervised learning (finding structure in unlabeled data), and reinforcement learning (learning via feedback/rewards). In the thesis, machine learning (ML) is the broader category encompassing methods like deep learning neural networks that have been used for tasks such as classifying ECG signals or predicting health outcomes.

Multimodal: In AI, "multimodal" refers to the capability to process and integrate multiple types of data (modalities) at once, such as text, images, audio, and signals. A multimodal AI model could, for instance, take both an ECG waveform image and a written patient history as inputs and combine that information to make a diagnosis. The thesis mentions multimodal systems in the context of advanced models like GPT-4V and Gemini, which aim to handle both language and vision. The promise of multimodal AI in healthcare is that it can consider varied data (e.g., ECG tracings, lab results, clinical notes) together, potentially providing more comprehensive clinical decision support.

Myocardial infarction: Commonly known as a heart attack, it is the injury or death of a portion of the heart muscle (myocardium) due to prolonged ischemia (lack of blood supply). Myocardial infarction (MI) typically occurs when a coronary artery is blocked, and it is often recognized on an ECG by specific changes such as ST-segment elevation, T-wave inversion, and the development of Q waves (depending on the type of MI). The thesis discusses the importance of quickly identifying signs of myocardial infarction on ECG (e.g., ST-elevations in a STEMI, which is a ST-elevation MI) because timely reperfusion therapy (restoring blood flow, e.g., via angioplasty or thrombolysis) can be lifesaving.

Neural network: A computational model inspired by the human brain's network of neurons. Neural networks consist of layers of interconnected "neurons" (or nodes) that process input data and can learn to perform tasks by adjusting the strengths (weights) of these connections. When such networks have many layers, they are called deep neural networks (the basis of deep learning). Neural networks are powerful for recognizing complex patterns and have been used in the thesis context for interpreting ECG signals and images. For example, an appropriately

trained neural network can classify different heart rhythms or detect conditions from ECG input by learning from many examples.

Overreliance: In the context of this thesis, overreliance refers to depending too much on the AI's suggestions without sufficient critical evaluation, which can degrade decision quality if the AI is wrong. Overreliance is closely related to automation bias; it's evident when clinicians accept AI output as true by default. The study observed that less experienced doctors were more prone to overreliance – for instance, they often followed the AI's incorrect advice, sometimes even changing correct answers to incorrect ones based on a misguided AI suggestion. Managing overreliance is critical, which is why training and trust calibration are emphasized as safeguards when introducing AI into clinical workflows.

p-value: A statistical metric that indicates the probability of obtaining the observed results (or something more extreme) if there were actually no true effect or difference (null hypothesis is true). A small p-value (conventionally < 0.05) suggests that the observed difference or association is unlikely to be due to chance alone and is considered statistically significant. For example, $p < 0.001$ in the results indicates an extremely strong evidence against the null hypothesis. In the thesis, p-values are reported for tests comparing diagnostic accuracies and behavior changes – a significant p-value means the differences (perhaps between using AI vs not using AI, or between groups of physicians) are real and not just random variation.

Pearson correlation: A statistical measure of linear relationship between two continuous variables, giving a coefficient (r) between -1 and 1. An r of 1 means a perfect positive linear correlation, -1 a perfect negative correlation, and 0 means no linear correlation. Pearson's correlation assumes data are roughly normally distributed. In the study, Pearson's r is used for correlations such as between years of experience and number of answer changes (when those variables meet assumptions). For instance, a negative Pearson correlation ($r \approx -0.41$) between AI familiarity and answer changes suggests that as familiarity with AI increases, the tendency to change answers (potentially due to AI influence) decreases.

Phi coefficient: A measure of association for two binary variables, often used as an effect size for 2x2 chi-square tests. The phi (ϕ) coefficient ranges from 0 (no association) to 1 (perfect association) in absolute value, with higher values indicating a stronger relationship. In this thesis, ϕ is reported alongside chi-square tests to indicate the strength of group differences (for example, $\phi \sim 0.8–0.9$ signifying very strong differences in outcomes between groups). If one treats a diagnostic outcome (correct/incorrect) and group (e.g., AI vs no AI, or different physician levels) as binary factors, phi provides an interpretable effect size of how tightly those are related ($\phi > 0.5$ was considered a large effect here).

Power analysis: A procedure in experimental design used to determine the sample size needed for a study to reliably detect an effect of a given size. It involves the interrelationship of statistical power (typically set at 0.8 or 0.9 for 80–90% power), effect size, significance level (α, usually 0.05), and sample size. A power analysis asks, for example: "Given the expected difference in accuracy with and without AI and the variability, how many participants do we need to have a high chance of finding a statistically significant result?" In this thesis, a priori power analysis was done using G*Power, which justified enrolling 25 physicians to achieve >95% power to detect

the differences of interest (meaning the study is very unlikely to miss a true effect of AI assistance if it exists).

Reperfusion therapy: Treatments aimed at restoring blood flow to an area that has suffered ischemia. In the context of myocardial infarction (heart attack), reperfusion therapy can include thrombolytic (clot-busting) drugs or mechanical interventions like percutaneous coronary intervention (angioplasty with stent placement) to reopen blocked arteries. The thesis mentions that timely identification of conditions like an ST-elevation MI on an ECG is critical because it triggers urgent reperfusion therapy, which can salvage heart muscle and significantly improve outcomes if done promptly.

Retrieval-Augmented Generation: Often abbreviated RAG, this is a technique in which a language model is combined with an external knowledge retrieval process. Before generating an answer, the model fetches relevant information (e.g., from a database or documents) and uses it to inform its response. The result is that the AI can provide not only answers but also cite sources or evidence, increasing transparency. In the thesis, RAG is mentioned as a way to address the AI "black box" problem by letting the model retrieve and show supporting information for its conclusions (for instance, giving references or data points when suggesting a diagnosis), thereby making its output more explainable and trustworthy.

Sensitivity and specificity: Metrics used to evaluate diagnostic test performance. Sensitivity (true positive rate) is the percentage of actual positive cases correctly identified by the test – for example, the proportion of patients with a disease that the AI correctly flags as having the disease. High sensitivity means few false negatives. Specificity (true negative rate) is the percentage of actual negative cases correctly identified the proportion of healthy individuals the test correctly identifies as not having the disease. High specificity means few false positives. In the context of AI ECG interpretation, a model might achieve sensitivity and specificity comparable to cardiologists for detecting certain arrhythmias. An ideal diagnostic tool has both high sensitivity (catches most of those who are ill) and high specificity (doesn't falsely label healthy people as sick).

Shapiro–Wilk test: A statistical test to assess whether a set of data is normally distributed. It is often used before parametric analyses (like t-tests or ANOVA) to check the normality assumption. The test provides a W statistic and a p-value; a non-significant p-value ($p > 0.05$) suggests the data do not significantly deviate from normality (i.e., they are roughly normal), whereas a significant result indicates deviation from normal distribution. In the thesis, Shapiro–Wilk tests were performed on accuracy data for each group of physicians. For instance, if the less experienced group's scores violated normality (as noted by a $p = 0.042 < 0.05$), a non-parametric test (Kruskal–Wallis) was chosen for comparisons involving that group's data.

Spearman's rank correlation: A non-parametric measure of correlation between two ranked (ordinal) or continuous variables. It assesses how well the relationship between two variables can be described by a monotonic function. Spearman's correlation coefficient (ρ, "rho") ranges from -1 to 1, similar to Pearson's r, but it is used when data do not meet Pearson's assumptions (e.g., non-normal distribution or ordinal data). In the thesis, Spearman's ρ was used for correlations involving variables like years of experience or familiarity with AI (especially when distributions were skewed). For example, a Spearman $\rho = -0.72$ between clinical experience and

magnitude of improvement with AI (noted as a strong negative trend) would imply that as experience increases, the relative boost one gets from AI tends to decrease (though in the study this particular correlation did not reach statistical significance).

Statistical power: The probability that a study will detect an effect or difference when there is one truly present (i.e., the test correctly rejects a false null hypothesis). A higher power means a lower chance of a Type II error (missing a true effect). Power is influenced by sample size, effect size, variability, and significance threshold. For instance, a study with 96% power (as reported for key comparisons in this thesis) is very likely to detect meaningful differences in diagnostic accuracy due to AI assistance if those differences exist. Researchers typically aim for at least 80% power in study design. Achieving high power in this thesis ensures confidence that if AI truly improves diagnostic performance (or affects it), the study would pick it up.

Statistical significance: An indication that the result observed is unlikely to be due to chance alone, according to a predefined threshold (significance level, α). When a result is "statistically significant," the p-value is below α (commonly 0.05). This suggests that the finding (e.g., an improvement in accuracy with AI, or a difference between groups) is real with a high level of confidence. It does not measure the size or importance of the effect, just that it is unlikely to be a random fluke. In the thesis, phrases like "statistically significant improvement" mean that with AI support, physicians' accuracy increased in a way that passed the significance test ($p < 0.05$), reinforcing that the AI's impact is verifiable and not just due to random variation in this sample.

Supraventricular tachycardia: Often abbreviated SVT, it is a rapid heart rhythm originating above the ventricles (in the atria or AV node). SVT typically presents as a sudden onset of a fast heartbeat that can be regular (often 150-250 beats per minute). On an ECG, SVT shows a narrow QRS complex tachycardia (unless there is aberrancy) with absent or retrograde P waves. In the discussion, a case of wide-QRS tachyarrhythmia was noted cardiologists assumed it was ventricular tachycardia (a dangerous rhythm from the ventricles), whereas GPT-4o identified it as an SVT with aberrancy (an atrial rhythm appearing with a wide QRS due to abnormal conduction). Distinguishing SVT from ventricular tachycardia is critical, as treatments differ, and it was an example of AI potentially catching what humans missed or vice versa.

Swarm learning: A decentralized machine learning approach related to federated learning, where insights from multiple local datasets are combined without central data pooling. In swarm learning, edge devices or institutions train models locally on their data and then share only learned parameters or weight updates (often using blockchain or secure consensus mechanisms to aggregate them). The term "swarm" implies multiple agents learning cooperatively like a swarm of bees. This method enhances privacy and robustness, as no single central server holds all data. The thesis references swarm learning as one of the techniques to ensure AI models (especially in medicine) can learn from widespread data (e.g., from different hospitals) without violating patient privacy or data governance rules.

Tamhane's T2 test: A post-hoc statistical test used after an ANOVA when comparing group means, especially when the assumption of equal variances is violated. Tamhane's T2 is a conservative method that does not assume equal variances between groups and is used to determine which specific groups differ from each other. In the thesis, after finding a significant difference across physician groups (with ANOVA or Kruskal–Wallis), post-hoc tests like Tukey's

HSD or Tamhane's T2 were applied. For instance, Tamhane's T2 was used when variances were unequal, to confirm, say, that each physician group's accuracy was significantly different from the others.

Transformer: A modern neural network architecture that has revolutionized natural language processing and is also applied in other domains. Transformers use a mechanism called self-attention to weigh the importance of different parts of the input sequence, enabling them to capture context over long ranges more effectively than previous recurrent neural networks. Large language models like GPT-3 and GPT-4 are built on Transformer architectures. The thesis indirectly references transformers when discussing LLMs (since GPT models use them). Transformers allow models to handle language (or even combined modalities) with unprecedented scale and coherence, which is why LLMs can generate fluid text and reason over complex input.

Triage: In medicine, triage is the process of prioritizing patients or cases based on the urgency of their condition, ensuring that those who need immediate attention get it first. In the context of AI, an AI triage system might automatically flag critical cases (for example, an AI reading continuous ECG telemetry might alert a doctor to a potentially fatal arrhythmia immediately). The thesis mentions using AI in triage, such as AI co-pilots monitoring data and alerting human teams to high-risk findings. By acting as a triage tool, AI can help manage workflow by drawing focus to the most pressing issues among many, e.g., highlighting an ECG that shows an acute problem out of a batch of normal ones.

Trust calibration: The process of aligning the user's trust in an AI system with the system's actual reliability and uncertainty. Proper trust calibration means the clinician trusts the AI when it has proven accurate or provides confident, well-founded recommendations, but remains cautious or double-checks when the AI is less certain or has potential to be wrong. The thesis emphasizes trust calibration as essential users should not blindly trust AI (overreliance) nor ignore it altogether but rather calibrate their trust. This can be facilitated by AI systems communicating their confidence levels or rationale, and by training clinicians to understand AI limitations. Effective trust calibration ensures that clinicians benefit from AI support when appropriate while maintaining healthy skepticism to catch AI's mistakes.

Tukey's HSD test: Short for "Tukey's Honestly Significant Difference" test, this is a post-hoc analysis used after ANOVA when you find a significant overall difference and want to know which specific group means differ. Tukey's HSD assumes equal variances and is good for pairwise comparisons while controlling the overall Type I error rate. In the study, Tukey's HSD was used when comparing multiple physician groups' performances under conditions where ANOVA showed a difference (and assumptions were met). For example, it helped confirm that cardiologists performed significantly better than both internist groups, and that experienced internists outperformed less experienced ones, by providing adjusted p-values for each pair of groups.

Ventricular tachycardia: Often called VT, it is a fast heart rhythm originating in the ventricles (the lower chambers of the heart). VT is a potentially life-threatening arrhythmia because it can compromise blood circulation or degenerate into ventricular fibrillation (which is fatal if not treated immediately). On an ECG, ventricular tachycardia typically appears as a series of wide

QRS complexes at a high rate (often 120-250 beats per minute) and no preceding P waves. It's mentioned in the thesis discussion as part of a challenging ECG case (wide-QRS tachyarrhythmia) where distinguishing VT from an SVT with aberrancy was critical. Identifying VT correctly is important for prompt treatment, and it was an example used to illustrate differences in AI vs human interpretation on tough cases.

ABBREVIATIONS

TERM	EXPLANATION
AF	Atrial Fibrillation
AI	Artificial Intelligence
ANOVA	Analysis of Variance
API	Application Programming Interface
AUC	Area Under the Curve
AV	Atrioventricular
CI	Confidence Interval
ECG	Electrocardiogram
EHR	EHR Electronic Health Record
F	F statistic
GPT	Generative Pre-trained Transformer
GPT-4V	A variant of GPT-4 that includes vision capabilities
GPT-4o	A specific tailored version of GPT-4
H	Kruskal–Wallis H statistic
LLM	Large Language Model
LSTM	Long Short-Term Memory
ML	Machine Learning
NTUA	National Technical University of Athens
o1 Mini - Preview	Preview OpenAI model
QRS: QRS complex	Ventricular depolarisation spike on an ECG
RAG	Retrieval-Augmented Generation
R.E.D.C	Research Ethics and Deontology Committee
ROC	Receiver Operating Characteristic
r	Pearson's correlation coefficient
SPSS	Statistical Package for the Social Sciences
STEMI	ST-Elevation Myocardial Infarction
SVT	Supraventricular Tachycardia
t	Student's t statistic
USMLE	United States Medical Licensing Examination
VT	Ventricular Tachycardia
Z	Standardised Z statistic
χ^2	Chi-square statistic
ϕ	Phi coefficient
η^2	Partial eta squared
ρ	Spearman's rho

1 INTRODUCTION

1.1 AI IN HEALTHCARE AND CLINICAL DIAGNOSTICS

Artificial intelligence (AI) has become a key driver of innovation in healthcare, especially in clinical diagnostics. Over the past decade, AI models particularly deep learning systems have reached impressive levels of accuracy in interpreting medical data, including images and physiological signals.

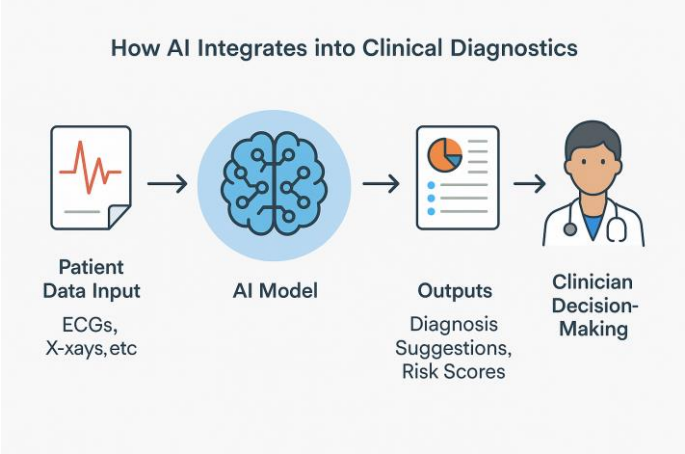


Figure 1 How AI Integrates into Clinical Diagnostics

In some cases, AI can detect subtle diagnostic patterns that clinicians might miss. For example, AI-based tools have been able to detect cardiac arrhythmia on electrocardiograms (ECGs) with up to 99% accuracy, even outperforming physicians on some rhythm classification tasks. These successes show how AI can support clinical decision-making by serving as a powerful pattern recognition tool.

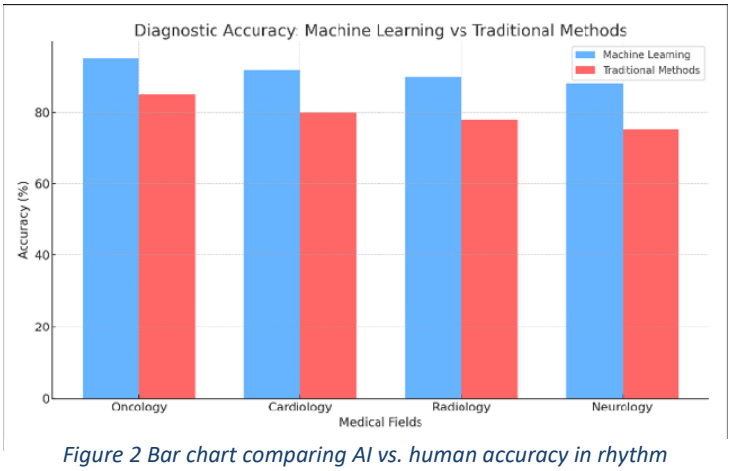


Figure 2 Bar chart comparing AI vs. human accuracy in rhythm classification tasks
(https://www.researchgate.net/figure/Comparative-Bar-Chart-Showing-the-Diagnostic-Accuracy-of-Machine-Learning-Versus_fig1_385395492)

Rather than replacing clinicians, well-designed AI systems are increasingly seen as assistive partners that can handle routine tasks or flag abnormalities, allowing physicians to focus on more complex decision-making. Many experts agree that AI should enhance not replace human intelligence in medicine, helping improve efficiency while the physician remains responsible for the final decisions[1]. This paradigm has led to the "AI co-pilot" concept in diagnostics, where AI supports clinicians with interpretation and decision-making.

1.2 Large Language Models in Medicine: Evolution, Promise, and Challenges

Among the newest and most influential AI tools in medicine are large language models (LLMs). These systems like OpenAI's GPT series are trained on massive amounts of text, allowing them to generate human-like language and reason through complex information. LLMs have quickly gained attention in healthcare for their ability to understand and produce text, combine different types of data, and even handle multi-modal reasoning[2]. Modern LLMs can pull together medical knowledge to answer questions, draft clinical notes, and provide explanations skills that show real promise for clinical decision support.

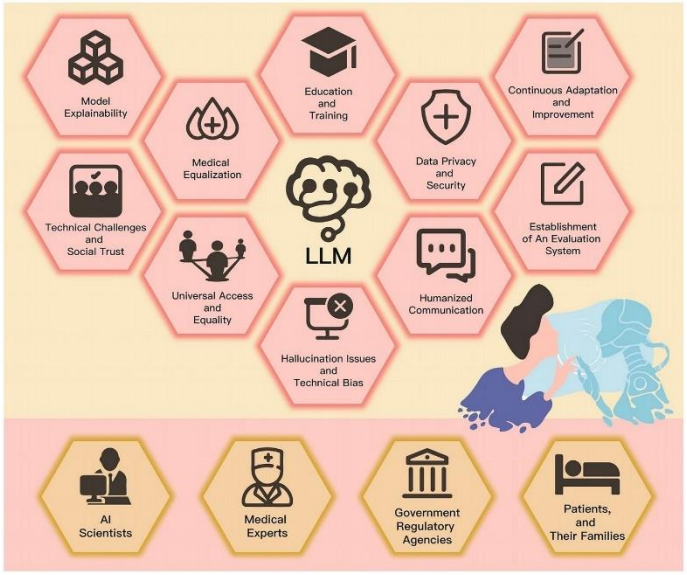


Figure 3 Capabilities of LLMs in Healthcare"
(https://blog.gopenai.com/the-future-of-medicine-exploring-the-potential-and-challenges-of-llms-in-healthcare-29aac7944e67)

Notably, models like GPT-4 have performed surprisingly well on medical competency exams, even approaching or surpassing the passing threshold for United States Medical Licensing Exam (USMLE) questions[3]. LLMs have also been tested as tools to help reduce clinicians' documentation workload for example, by generating discharge summaries and offering instant access to medical knowledge for learning purposes[3]. These advances suggest that LLMs could become versatile assistants in both clinical practice and medical education, thanks to their ability to understand context and generate reliable responses.

However, using LLMs in medicine also comes with important challenges. First, these models don't have real clinical experience or reasoning based on real-world practice they only simulate reasoning by learning patterns from text. As a result, they can produce answers that sound confident and convincing but are factually wrong or not clinically validated a well-known problem called AI "hallucination." On top of that, most general LLMs at the time of writing can't process visual data like medical images or waveforms without special modifications [2], [3]. This is a major limitation in fields like radiology or cardiology, where being able to interpret visual or spatial signals is essential. LLMs also face trade-offs when it comes to domain specificity their broad training can sometimes make them less accurate with highly specialized medical details[2]. Fine-tuning these models for medical use requires significant computational resources, and many of the most advanced models are proprietary, which creates barriers to access (researchgate.net). Beyond the technical challenges, there are also ethical and safety concerns. LLMs can accidentally reveal private patient information or reflect biases present in their training data. Their lack of transparency often called the "black box" problem can make it harder for clinicians to trust the AI, especially when the reasoning behind its suggestions isn't clear. Reliability is also crucial, if an AI occasionally gives unsafe advice, it can quickly undermine clinicians' confidence in the tool. While large language models hold tremendous potential as clinical assistants, unlocking that potential will require addressing challenges around accuracy, transparency, bias, and integration into clinical workflows [4], [5]**.** Researchers emphasize the need for rigorous validation, better multimodal capabilities, and human-in-the-loop designs to ensure that LLMs function as safe, reliable tools that support expert care rather than acting as unchecked or unreliable sources of information [3], [4], [5]

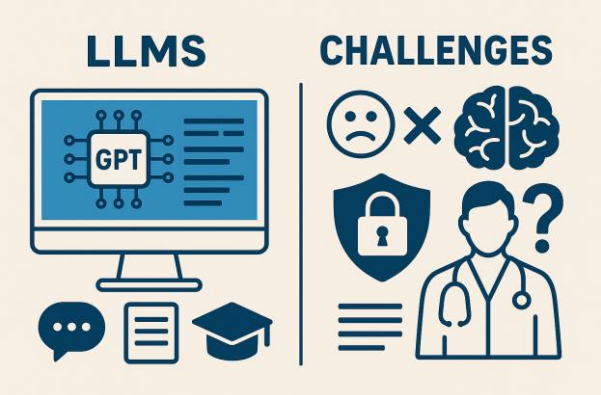


Figure 4 Challenges of Using LLMs in Clinical Settings

1.3 ECG Interpretation in Clinical Workflows and the Expertise Gap

In clinical practice, few skills are as common yet as deceptively complex as interpreting an electrocardiogram (ECG). The 12-lead ECG is a first-line diagnostic tool for a wide range of heart conditions, from arrhythmias to ischemia, and it plays a crucial role in emergency and acute care. In fact, a significant portion of acute medical visits involve cardiac evaluation about 20% of emergency department cases present with cardiovascular symptoms [2], [3], [4], [5], and the ECG often guides both triage and treatment decisions.

Fast and accurate ECG interpretation can be life-saving for example, by spotting an evolving myocardial infarction in time to avoid delays in reperfusion therapy. On the other hand, misreading an ECG can result in missed diagnoses or inappropriate treatments, both of which can have serious consequences [4], [5], [6]. Despite its importance, mastering ECG interpretation is difficult. It requires learning to recognize a wide range of normal variations and abnormal patterns, and not all clinicians receive the same level of training in this skill[4], [5]. As a result, accuracy in ECG interpretation varies widely between clinicians and is closely linked to their experience and specialty. A systematic review found that practicing physicians, on average, achieve only about 68% accuracy in ECG interpretation, while cardiology specialists perform somewhat better, averaging around 75% accuracy [7], [8]. Less experienced doctors, such as junior residents, often perform even worse on challenging ECGs, highlighting a competency gap that can directly impact patient care. One study found that up to one-third of ECG interpretations by general physicians contained significant errors, and about 11% of those errors led to inappropriate patient management [9]. Common mistakes include missing subtle signs of ischemia, failing to recognize uncommon arrhythmia, or misinterpreting pacemaker rhythms all of which highlight how complex ECG interpretation really is.

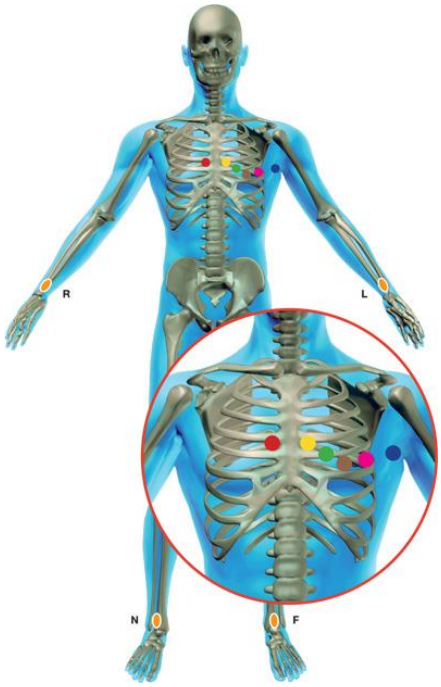


Figure 6 Illustration of a standard 12-lead ECG placement on a human torso
(https://www.numed.co.uk/news/12-lead-ecg-lead-placement-guide)

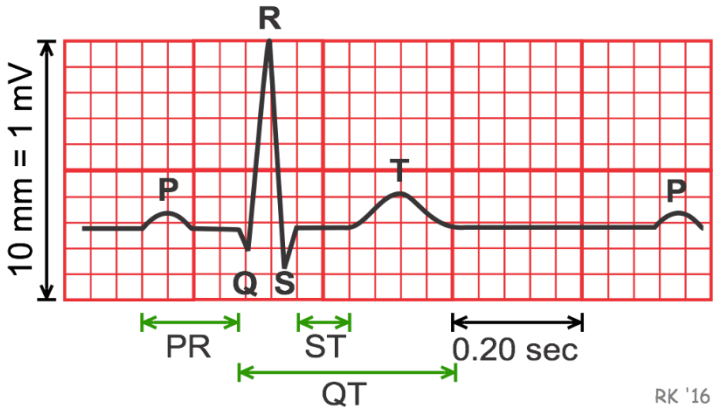


Figure 5 Standard ECG Interpretation
(https://www.cyberdefinitions.com/definitions/ECG.html)

The challenges less experienced physicians face with ECG interpretation are well documented. Unlike experienced cardiologists who have reviewed thousands of tracings, trainees and non-cardiology doctors often struggle with pattern recognition and may lack confidence in their interpretations. They may correctly spot obvious abnormalities like clear ST-elevation in a myocardial infarction but miss more subtle signs or mistakenly label harmless findings as pathological. This variability can lead to suboptimal patient outcomes and also contributes to

a well-recognized anxiety among trainees often referred to as "ECGphobia."[10]. Traditionally, the solution to this competence gap has been more training and hands-on experience. However, persistent error rates along with the lack of a standardized, highly effective method for teaching ECG interpretation remain ongoing challenges[9], [10], there is growing interest in tools that can assist clinicians during the interpretation process. AI-powered ECG analysis has emerged in recent years as a potential solution to this need. Dedicated machine learning algorithms using computer vision on ECG waveforms or signal analysis have demonstrated the ability to detect certain arrhythmias or conditions with expert-level accuracy. Until recently, these algorithms were highly task-specific for example, designed solely to detect atrial fibrillation or hypertrophy and mainly functioned as automated ECG readers that generated computer-based diagnoses on the ECG printout. These tools act as a "second pair of eyes," but an inflexible one, since they can't explain their reasoning or adapt to unusual cases. The question now is whether a more general AI specifically, a large language model with broad medical knowledge could serve as a smarter diagnostic assistant for ECG interpretation [11]. Such an AI co-pilot wouldn't just suggest a diagnosis but could also explain its reasoning or answer questions, potentially offering a more interactive and intuitive form of support than a traditional, static algorithm.

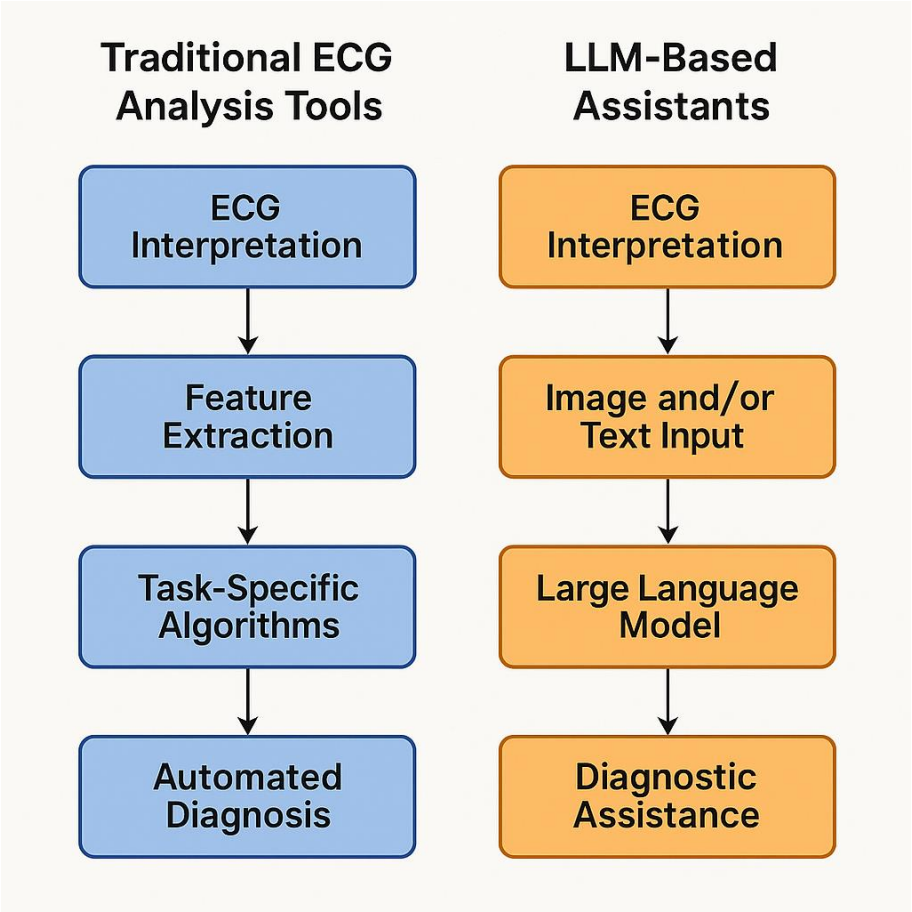


Figure 7 Comparing traditional ECG analysis tools vs. LLM-based assistants

1.4 Research Problem Statement

This thesis addresses the core research problem of evaluating a large language model's role as a diagnostic co-pilot in ECG interpretation, across physicians with varying levels of expertise, and understanding the risks associated with overreliance on AI assistance**.** In particular, we focus on GPT-4o, a state-of-the-art LLM based on GPT-4, and investigate how its suggestions might influence and improve clinicians' reading of ECGs [12], [13], [14]. The motivation for this research comes from two key insights discussed earlier: (1) Less experienced physicians often struggle with accurate ECG interpretation, and even experienced clinicians have limitations suggesting that decision support in this area could be highly valuable; and (2) LLMs like GPT-4 have reached a level of sophistication where they may be able to offer relevant, context-aware diagnostic advice, but their real-world effectiveness and safety in this role are still unproven. Recent early studies have suggested the potential of LLMs in ECG interpretation. For instance, one investigation of ChatGPT (based on GPT-3.5/4) found that its interpretations often aligned with cardiologists on many ECG features, but with notable gaps in critical areas, such as detecting subtle ECG changes linked to impending adverse events[2], [15]. In that study, ChatGPT tended to overestimate risk, flagging more cases for major cardiac events than the human experts did[15], [16]. This highlights a key concern: Will an AI's strengths like pattern recognition and consistency genuinely help clinicians, or could its weaknesses such as false alarms or missed subtle findings end up misleading them? The balance between benefit and risk is still uncertain.

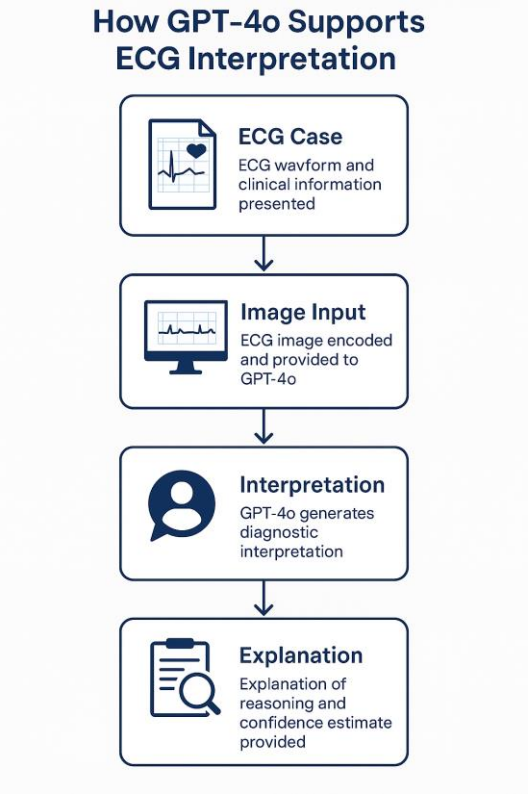


Figure 8 How GPT-4o Supports ECG Interpretation

GPT-4o is one of the most advanced models available, and unlike earlier AI systems that were limited to pre-programmed ECG algorithms, GPT-4o can process case information including ECG waveforms via image input or text-based descriptions of ECG findings and generate a diagnostic interpretation with an explanation. This flexibility could make it a powerful, interactive tool for clinicians interpreting ECGs. However, before this study, there had been no formal evaluation of GPT-4o's performance or its impact in a clinician-in-the-loop ECG interpretation setting. We still don't know how accurate GPT-4o is across a wide range of ECG cases, or how clinicians might use or potentially misuse its advice in real practice. A key part of the problem is understanding overreliance: if the AI provides an incorrect suggestion, will physicians especially those less confident in ECG interpretation defer to it and make an error they otherwise wouldn't have? Automation bias has been documented in other human-AI

decision-making studies, showing that even skilled professionals can be influenced by an AI suggestion and change a correct answer to an incorrect one [6], [9], [17], [18]. These kinds of false "corrections" can be damaging. This leads to the core research question: Can an advanced LLM like GPT-4o meaningfully improve physicians' accuracy and confidence in ECG interpretation as a diagnostic co-pilot and what are the risks, especially the risk of clinicians being misled by incorrect AI suggestions? Answering this requires not only measuring diagnostic performance with and without AI support but also understanding how physicians at different experience levels interact with the AI who benefits the most, who might be negatively impacted, and under what circumstances. By systematically exploring these questions, this thesis aims to provide an evidence-based assessment of GPT-4o's usefulness as a clinical tool for ECG interpretation, while also contributing to the broader understanding of how human-AI collaboration works in medical diagnosis [19].

1.5 Objectives and Research Questions

Building on the problem statement, this study aims to evaluate how GPT-4o impacts ECG interpretation and to examine how trust, accuracy, and error occur when physicians collaborate with the AI. The research is guided by several key questions:

1. **Diagnostic Performance Improvement:** How much does GPT-4o assistance improve physicians' accuracy in interpreting ECGs? This study will measure each physician's performance before and after receiving AI suggestions to determine whether GPT-4o leads to a meaningful improvement or not in diagnostic accuracy.
2. **Experience-Level Differences:** How does AI assistance impact different types of physicians with different levels of experience? This study examines whether the benefits or potential risks of GPT-4o's suggestions vary among expert cardiologists, experienced internists, and less experienced internists. An underlying hypothesis is that less experienced doctors may benefit more from AI support because they have more knowledge gaps, but could grow more dependent on AI.
3. **Case Difficulty and AI Utility:** Does GPT-4o's impact depend on how complex the ECG case is? We analyzed everyday and more challenging ECG cases separately to assess whether the AI provided greater support in detecting complex or uncommon patterns that typically challenge less experienced physicians, or whether its performance declined under increased diagnostic difficulty.
4. **AI Suggestion Accuracy and Physician Trust:** What happens when the AI's suggestion is either intentionally incorrect or simply a wrong prediction? This question focuses on the risk of overreliance. Specifically, we examine how often participants accept incorrect recommendations from GPT-4o and whether this varies based on experience level. Do junior physicians abandon their correct initial judgments because the AI suggests something different? Are senior physicians more likely to catch the AI's mistakes? By including scenarios with intentionally incorrect AI suggestions, this study aims to measure how often and under what conditions AI-induced diagnostic errors occur.

5. **Decision Change Analysis:** How do GPT-4o's suggestions influence physicians' decision-making process? We will track how often physicians change their initial diagnosis after seeing the AI's recommendation. Among those changes, we distinguish between beneficial changes when an incorrect answer is corrected with the help of the AI and harmful changes when a correct answer is switched to an incorrect one because of the AI. This analysis helps reveal how the AI impacts decision-making: whether it mainly acts as a safety net that catches errors or whether it sometimes leads physicians to override correct decisions.

6. **Physician Confidence and Workflow Perception:** While more difficult to measure, this study also explores how physicians perceive AI's influence. Do they feel more confident in their final answers when supported by GPT-4o, or do they experience any sense of reduced autonomy? Although this is not a primary outcome, we collect qualitative feedback to help contextualize the quantitative findings.

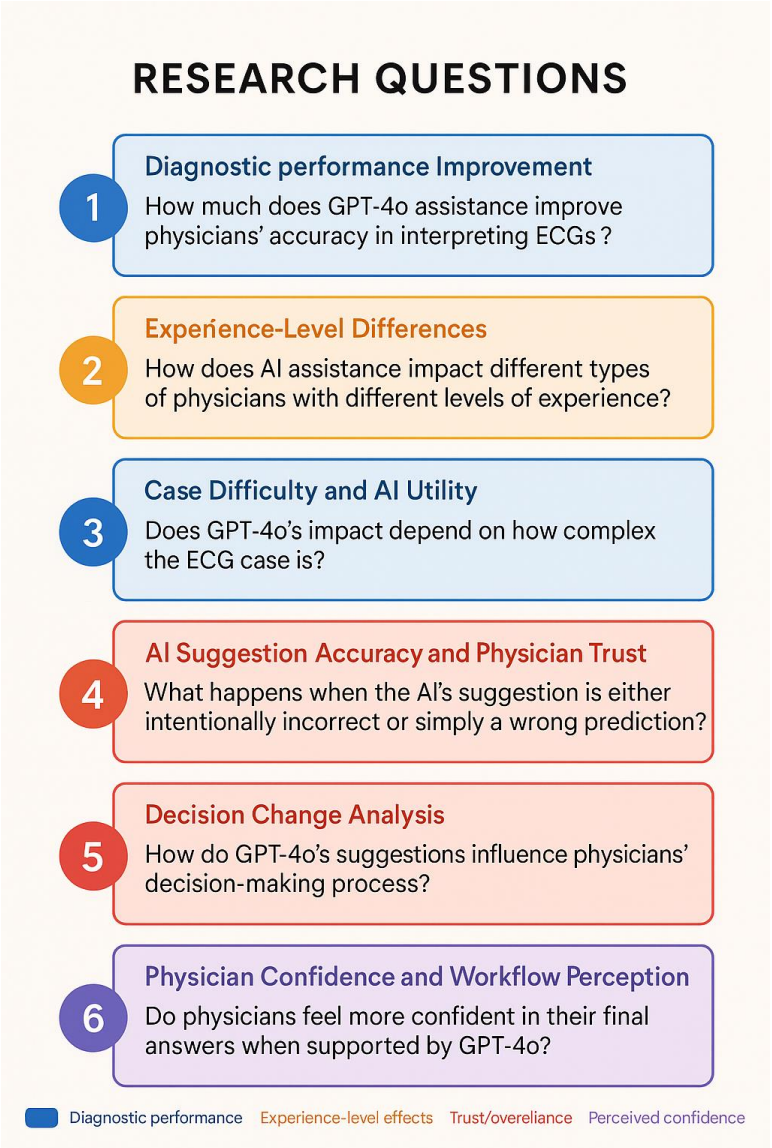


Figure 9 Infographic of the 6 research questions, color-coded by theme

1.6 Methodology Overview

To address the above questions, we designed a controlled experimental study involving 25 physicians and a standardized set of ECG interpretation cases. The participants were divided into three groups representing different levels of ECG expertise: 10 cardiologists (with 15–25 years of clinical experience, experts who routinely interpret ECGs), 10 experienced internal medicine physicians (with 15–25 years of clinical experience, likely comfortable with common ECG findings but not specialists), and 5 less-experienced internal medicine physicians (recent graduates with under 5 years of experience, expected to have the most difficulty with ECGs). This grouping allowed us to compare the impact of AI across different levels of expertise. All participants were recruited from hospital settings and gave informed consent, with ethical approval obtained for the study protocol.

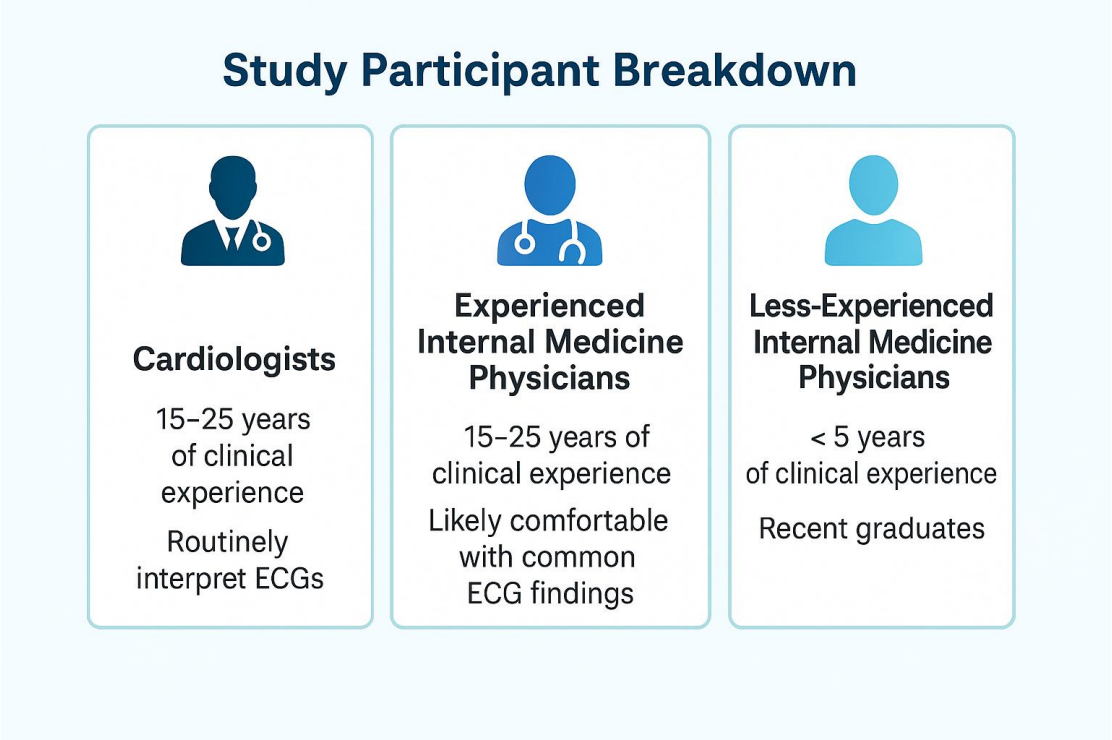


Figure 10 Study Participant Breakdown

The diagnostic task involved a questionnaire of 50 ECG cases, each presented as a multiple-choice question with five possible diagnoses. These cases were carefully selected from reputable cardiology case collections, primarily 150 ECG Cases by John R. Hampton [20], along with a cardiology question bank for additional challenging cases to ensure a broad range of clinical scenarios. We intentionally balanced the case mix as follows: 20 “everyday” ECG cases representing common, straightforward diagnoses such as typical atrial fibrillation or a classic ST-elevation myocardial infarction, and 20 “challenging” ECG cases representing more complex or atypical diagnoses such as uncommon arrhythmias, subtle signs of ischemia, or ECG changes related to electrolyte imbalances. This ensures that we could evaluate GPT-4o on both routine and difficult interpretations. In addition, a critical feature of our methodology was the inclusion of 10 special cases, drawn from the challenging category, in which the AI

intentionally provided an incorrect suggestion. These cases, designed in collaboration with a cardiologist, were crafted to be both realistic and deliberately misleading for example, an ECG showing pericarditis incorrectly labeled by the AI as a myocardial infarction. The goal of these cases was to assess how likely physicians are to be misled by incorrect AI advice, directly testing the risk of overreliance.

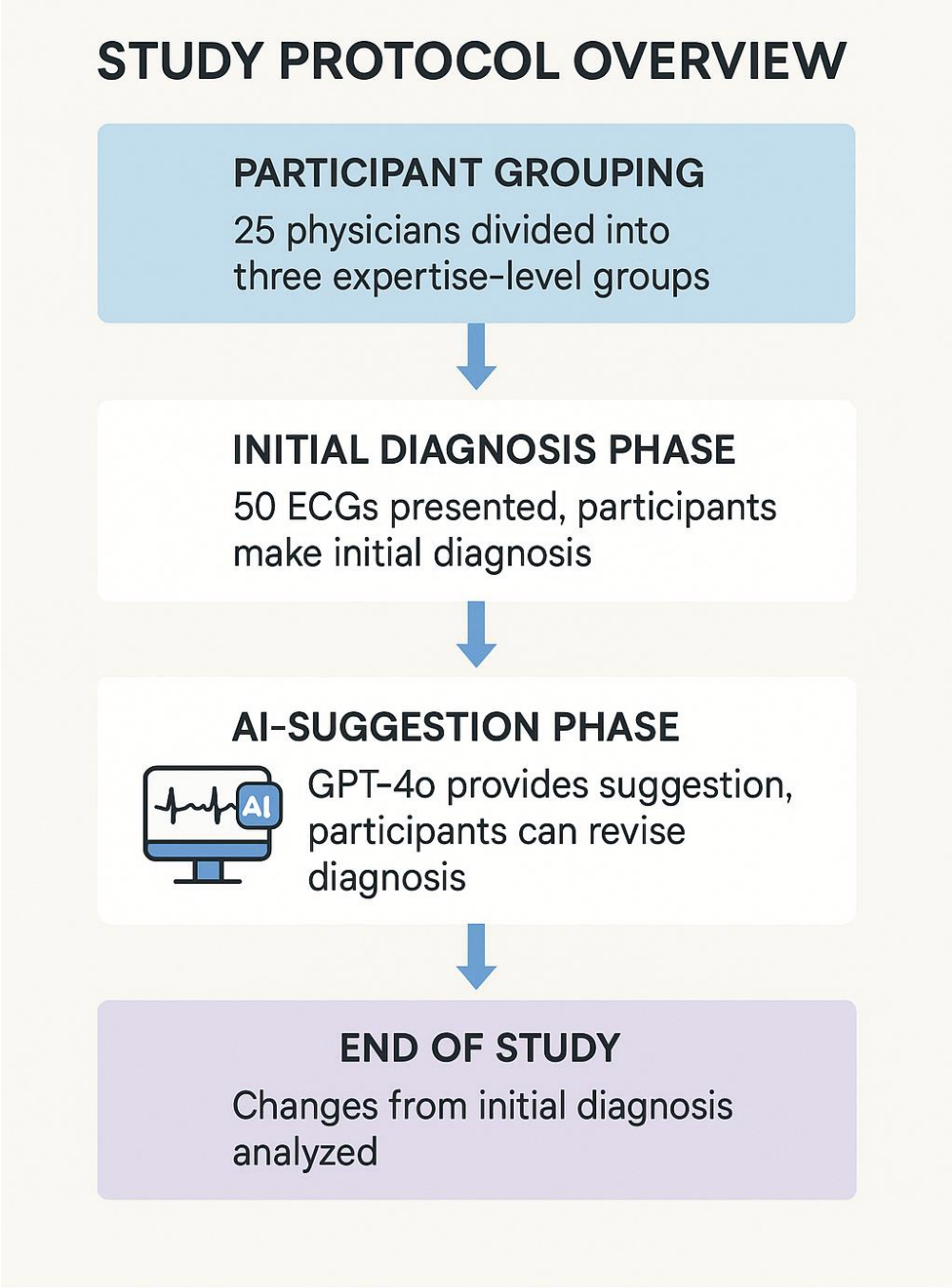


Figure 11 Study Protocol Overview

Each participant completed the 50-case ECG questionnaire in a two-phase process designed to isolate the impact of the AI intervention. In the Initial Diagnosis phase, physicians were shown each ECG along with relevant patient information such as age, sex, brief clinical context and selected their answer based solely on their own knowledge and judgment. They were unaware at this stage which cases were considered "difficult" or that any would have false AI suggestion, in order to avoid biasing their approach. In the AI-Suggestion phase, after the physician submitted their initial answer for a case, the same case was presented again this time accompanied by a diagnostic suggestion from GPT-4o. The AI's suggestion included the diagnosis it considered most likely the answer it would choose, along with a brief explanation of its reasoning and a confidence estimate. GPT-4o's output was designed to mimic how a human consultant might explain an ECG, providing insight into why a particular diagnosis was recommended. After reviewing the AI's suggestion, participants could either keep their original answer or revise it. This process was repeated for all cases. Essentially, each physician acted as their own control we captured their accuracy on each case *before* and *after* AI input, allowing paired comparisons of performance.

To generate GPT-4o's ECG interpretations consistently, we used the OpenAI API with carefully designed prompts. Each ECG image was encoded and provided to the model along with standardized instructions intended to simulate a clinical reasoning process. The model referred to as "GPT-4o" had been internally validated as the best-performing option among several GPT variants tested for this task including GPT-3.5 Turbo, GPT-4, GPT-4 Turbo, o1 Preview, and o1 Mini, with GPT-4o demonstrating the highest accuracy in pilot testing. We ran the model in deterministic mode with low randomness settings to ensure it produced consistent answers across repeated trials. For each case, the model was queried five times to confirm it consistently selected the same answer before presenting its suggestion effectively simulating a confident "AI opinion" for the physicians to consider. In the 10 cases with intentionally incorrect suggestions, we bypassed the model's actual answer and instead displayed a plausible but incorrect diagnosis, accompanied by a convincing rationale. This approach simulated a scenario where the AI makes a confident mistake. Participants were unaware of this manipulation, allowing us to observe their natural reactions to a wrong AI recommendation.

All participant responses, both initial and final answers were collected via SurveyMonkey and stored in the platform for analysis. Data analysis focused on several comparisons. We calculated each physician's accuracy before AI assistance, after AI assistance, and the difference between the two, then compared these changes across the three groups. We also analyzed performance based on case difficulty. Statistical tests including paired t-tests or non-parametric equivalents for within-subject comparisons, and ANOVA or Kruskal-Wallis tests for between-group comparisons were used to assess whether any observed improvements with AI were statistically significant. Importantly, we quantified the outcome of each decision change: how often the AI led to a change from incorrect to correct a beneficial change, from correct to incorrect a harmful change, or resulted in no change. We then compared these outcomes across experience levels and case types. Particular attention was given to the 10 deceptive cases to determine how often physicians accepted the AI's incorrect suggestion and whether certain groups such as less experienced physicians were more susceptible. By

analysing these patterns, our goal is to draw meaningful conclusions about GPT-4o's reliability and whether it is advisable to use it as a diagnostic aid in clinical practice.

In summary, our methodology offers a controlled yet realistic simulation of using an LLM as a diagnostic co-pilot for ECG interpretation. It enables us to measure objective outcomes such as improvements or declines in accuracy while also capturing behavioral factors like trust and decision changes. More extensive description is available in **Chapter 3**.

1.7 Thesis Structure

The remainder of this thesis is organized into five chapters, each covering a key aspect of the study and contributing to the overall foundation for our conclusions:

- **Chapter 2: Literature Review** – This chapter reviews the background and prior research relevant to our study. It begins with an overview of AI applications in cardiology and diagnostics, highlighting studies where machine learning has improved the detection of cardiac conditions such as AI in arrhythmia detection and medical imaging. The chapter then explores the rise of large language models (LLMs) in healthcare, summarizing current research on how models like GPT-3 and GPT-4 have been applied or evaluated in medical settings. This includes their performance on medical licensing exams, their use in clinical documentation, and early experiments testing their role as diagnostic assistants. Additionally, it examines well-documented challenges in the literature, including AI hallucinations, bias, and the risks of automation bias or overreliance in human-AI collaboration. Together, this review highlights the key knowledge gap our study seeks to address the lack of empirical data on how LLMs support diagnosis in real-world clinical tasks like ECG interpretation.
- **Chapter 3: Methods** – This chapter provides a detailed explanation of the study design and procedures. It describes how participants were recruited and grouped, how the ECG cases were selected and validated, and outlines the step-by-step experimental protocol described earlier. It also details how GPT-4o was configured and integrated into the questionnaire, including the prompt engineering techniques and internal validation steps used. The chapter further explains the statistical analysis plan, defining how performance improvements were measured, how answer changes were categorized, and which statistical tests were applied for hypothesis testing. Ethical considerations and any methodological limitations are also addressed in this chapter.
- **Chapter 4: Results** – This chapter presents the quantitative findings of the study. It begins with an overview of GPT-4o's standalone performance on the ECG cases showing how often the AI was correct on its own to provide context. The main results follow, detailing physician accuracy before and after AI assistance, broken down by experience group and case difficulty. The chapter also includes a detailed analysis of decision change patterns specifically, the proportions of beneficial versus harmful changes with particular focus on the intentionally incorrect AI suggestion cases. Statistical significance for all comparisons is reported. Additionally, any notable observations such as specific ECG diagnoses where the AI consistently provided

helpful guidance or, conversely, frequently misled participants are discussed. The purpose of this chapter is to objectively present the outcomes of physician-AI interactions, laying the groundwork for interpretation and discussion in the following chapter.

- **Chapter 5: Discussion** – This chapter interprets the results in relation to the research questions and existing literature. It evaluates when GPT-4o functioned as an effective diagnostic co-pilot especially for less experienced physicians and when it contributed to errors due to automation bias or overreliance. The discussion considers the broader implications for using LLMs in clinical diagnostics beyond ECGs, reflects on GPT-4o's technical strengths and limitations, and compares the findings to previous studies. It also acknowledges the limitations of the study and suggests how future research could address them.
- **Chapter 6: Conclusion and Future Work** – This chapter summarizes the key takeaways, emphasizing how GPT-4o improved diagnostic accuracy for less experienced clinicians while also posing risks when its suggestions were incorrect. It highlights the importance of safeguards like AI explainability and user training. Recommendations for future research include testing with larger groups, comparing generalist models like GPT-4o to specialized ECG AI, and studying how clinicians' trust in AI evolves with long-term use. The thesis concludes with reflections on the promise and responsibility of integrating AI into clinical practice to truly support patient care.

In summary, this introduction has outlined the background, motivation, and importance of evaluating LLMs as diagnostic assistants, specifically for ECG interpretation. The following chapters will present evidence and analysis addressing whether GPT-4o can effectively serve as a diagnostic co-pilot for physicians and how to balance the benefits and risks of integrating AI into clinical decision-making.

2 LITERATURE SURVEY

2.1 AI IN CLINICAL DIAGNOSTICS: EMERGENCE OF LLMS AND DECISION SUPPORT

Artificial intelligence (AI) has become an increasingly important part of innovation in clinical diagnostics, providing tools that support earlier detection and better decision-making across many areas of medicine. Early AI systems in healthcare were often based on supervised deep learning models, trained on large, labeled datasets to perform specific tasks. For example, in cardiovascular medicine, deep neural networks have shown improved accuracy in interpreting medical images and signals, often uncovering diagnostic details that human experts might miss. Specifically, neural networks used in cardiac diagnostics have detected subtle patterns in imaging or ECG data that are linked to disease risk, sometimes outperforming traditional physician interpretation. These results show that AI can enhance clinicians' diagnostic abilities by acting as powerful pattern recognition tools [21], [22], [23].

A major recent development in healthcare AI is the rise of foundation models and large language models (LLMs). Unlike task-specific algorithms, LLMs like GPT-4 can process and generate human-like text, allowing them to hold flexible conversations, interpret clinical notes, and even combine different types of medical data. Modern LLMs use transformer-based designs and large-scale pre-training to combine different types of data including patient records, imaging, genomics, and more to offer a more complete picture of a patient's condition. This ability to integrate multiple data sources supports the idea that AI could eventually help synthesize complex clinical information into meaningful risk predictions or accurate differential diagnoses. For example, one advanced model has been designed to combine electronic health records (EHRs), ECG signals, and imaging to better identify patients at risk of sudden cardiac death potentially spotting candidates for preventive treatments, like defibrillators, more accurately than current methods[24], [25], [26], [27]. This shows how LLM-based AI systems are evolving beyond simple diagnostic tools into full clinical decision support systems that can reason across a wide range of medical data.

This shift has led to the idea of AI acting as a "co-pilot" in clinical practice. Instead of working on its own, AI is now increasingly seen as a collaborative partner that supports clinicians. In practice, an AI co-pilot might help review patient data, draft clinical notes, or suggest possible diagnoses all while the physician remains responsible for the final decision. Notably, some healthcare institutions in the U.S. have already started using LLM-powered tools for tasks like clinical notetaking, reflecting growing excitement about their potential to improve efficiency and quality of care [28], [29], [30]. Early pilot studies are promising. One report found that an adapted LLM did a better job than medical experts at summarizing clinical notes from patient records, showing how useful these models can be for handling time-consuming information synthesis. Similarly, a recent study evaluating an LLM on 149 real patient case vignettes covering diagnosis, patient communication, and management planning found that the AI performed as well as or better than physicians in most of these cases [31], [32]. These results

suggest that when given a well-defined problem, advanced AI can deliver decisions at a physician level or act as a reliable second opinion in complex situations.

2.2 Human–AI Collaboration and Co-Pilot Frameworks in Practice

The growing use of human–AI collaboration in medicine reflects how AI recommendations are being integrated directly into clinical workflows. The term "co-pilot" refers to AI systems designed to assist with routine cognitive tasks, helping clinicians focus more on complex decision-making and patient care. In practice, AI co-pilot models are already being used in fields like radiology for example, triaging scans for review as well as in pathology and emergency medicine for patient triage. In one example, an AI-powered tool has been integrated into electronic health records (EHR) to automatically draft replies to patient messages, which physicians then review and edit [33], [34], [35]. This type of deployment suggests that, in the future, routine or data-heavy parts of clinical work could be handled by AI, allowing physicians and AI to work together more efficiently than either could on their own.

Crucially, the co-pilot model emphasizes that the final responsibility and decision-making remain with human clinicians. AI suggestions are meant to inform and support, not replace, the clinician's decision-making. When implemented thoughtfully, this synergy can improve outcomes. For example, Topol [36] envisions multimodal AI assistants acting as 24/7 "attending physicians" in critical care settings, continuously monitoring patient data (like ECG telemetry) to catch infrequent but life-threatening changes and promptly alerting the human team. In ECG monitoring specifically, an AI co-pilot could analyze waveforms in real time and flag subtle arrhythmias or ischemic changes that might develop between routine checks. Such a system could act like a tireless second pair of eyes, helping improve patient safety by reducing the risk of missed findings. Another prospective use of LLM-based co-pilots is in medical training and education [37]. Generative AI can serve as a virtual tutor or simulator, allowing medical trainees to practice interpretation skills and receive instant feedback. For example, an LLM that explains ECG findings and the reasoning behind them could help junior doctors strengthen their understanding, essentially serving as an "automated second reader" for challenging cases.

Despite these promising developments, real collaboration between humans and AI in healthcare is still at an early stage, and several challenges still limit the excitement. Testing in real clinical settings is still a major gap. A recent systematic review found that most studies evaluating LLMs have used proxy tasks, like answering board-style questions, instead of testing them in real patient care situations [38]. Only about 5% of published evaluations up to early 2024 involved live clinical data or workflows [39]. Most studies have focused on accuracy in narrow tasks, like question-answering, while paying much less attention to important factors such as fairness, bias, and reliability in real-world use. This suggests that while lab-based performance of AI is often impressive, its translation to routine practice with all the messy complexity of real patients is less certain. In fact, early results are mixed. Some researchers found that chatbot diagnosticians can give superficial or sometimes unsafe advice, while others reported diagnostic accuracy close to that of clinicians when looking only

at whether the answers were correct. This gap highlights how important the evaluation criteria are. An AI might get the "right answer" often enough to perform well on a test but still fail to provide the context, judgment, or depth of explanation that a physician would offer in real patient care.

The current limitations of LLM-based AI in clinical decision-making are well documented in the literature. One major concern is AI "hallucinations," where the model generates information that sounds correct but is actually wrong. Without strong safeguards, an LLM can sometimes make up clinical facts or misstate guidelines, which could mislead users if not noticed. These models can also carry over biases from the data they were trained on. As Quer and Topol [36] note, an LLM can unintentionally reflect societal or racial biases found in healthcare data, which could lead to unfair or unequal recommendations. Addressing this requires careful checks for bias and possibly fine-tuning the model using more carefully selected datasets. Another issue is the lack of explainability. Clinicians may be skeptical to trust AI advice if the reasoning behind it is unclear the so-called "black box" problem. One possible solution being explored is retrieval-augmented generation (RAG), which allows LLMs to provide citations or evidence for their answers to make the process more transparent. Data privacy is another important challenge [40], [41]. Since LLMs are trained in large amounts of text, they could sometimes accidentally generate sensitive information, raising concerns about patient confidentiality. Techniques like federated learning or swarm learning, which train AI on distributed data without collecting it in one place, and differential privacy are being explored to help reduce these risks. In summary, while LLMs offer huge potential as clinical co-pilots bringing better efficiency, consistency, and possibly even higher quality of care the medical community is moving forward carefully. Strong testing in different clinical settings, along with safeguards to ensure accuracy, prevent bias, and protect privacy, are widely seen as essential before LLMs can be fully trusted as partners in patient care.

2.3 Applications of AI in Electrocardiogram Interpretation

Electrocardiogram (ECG) interpretation is a domain that has seen active exploration of AI tools, from traditional algorithmic analyzers to cutting-edge deep learning and LLM approaches. Because ECG reading is a routine part of clinical practice but still depends heavily on expert judgment, finding ways to automate or support this task with AI has been a long-standing goal in medical research. In fact, basic computer-based ECG interpretation systems have been embedded in ECG machines for decades to provide preliminary diagnoses. However, the accuracy of these systems has been limited, often requiring physicians to over-read to correct errors. This limitation created the opportunity for improved methods, and in recent years machine learning (ML) techniques, particularly deep learning, have dramatically advanced the field.

Early successes were reported with deep neural networks trained on large ECG datasets to detect cardiac arrhythmias. Notably, Hannun et al. [42] developed a neural network for classifying arrhythmias from single-lead ECG recordings that achieved cardiologist-level performance. The system could diagnose 12 types of cardiac rhythms with high sensitivity, matching experts in identifying conditions such as atrial fibrillation (AF) vs. normal rhythm.

Subsequent studies extended these approaches to the standard 12-lead ECG used in clinical practice. Chang et al. [43]for example applied a long short-term memory (LSTM) deep learning model to a large set of 12-lead ECGs and obtained remarkable results: the model's accuracy in classifying each of 12 rhythm diagnoses was ≥98%, with area-under-curve (AUC) metrics around 0.99 across the board. This near-perfect performance shows how far AI has come in recognizing patterns in ECG waveforms. It suggests that for certain well-defined tasks, like rhythm classification, a well-trained algorithm can match the consistency and accuracy of human specialists. In fact, there are studies showing cases where AI-based ECG interpretation outperformed physicians: One study found that an AI system outperformed not only general practitioners but also cardiologists in classifying arrhythmias [42]. Beyond arrhythmias, deep learning models have been developed to detect other conditions from ECGs that can be subtle or hidden, such as predicting left ventricular dysfunction, detecting hypertrophic cardiomyopathy, or even identifying patients at risk for pulmonary hypertension based on waveform patterns. These results highlight that ECG signals contain far more information than what is typically visible, and AI can help uncover complex patterns that connect those signals to clinical diagnoses or outcomes.

2.4 Benchmarking AI and LLMs vs. Physicians in ECG Tasks

With the rise of general-purpose AI like LLMs, researchers have started testing how well these models perform on ECG interpretation compared to human experts. A key question is whether modern AI can not only read waveforms but also mimic the way clinicians think when making diagnostic decisions from an ECG. Several recent studies have directly compared AI models, including LLM-based systems, to physicians on standardized ECG tasks.

One such study compared the latest GPT-4 model an LLM by OpenAI against cardiologists and emergency medicine doctors both with less than 5 years of experience after their residency in interpreting ECG cases. Investigators constructed a test of 40 ECGs split into 20 routine cases and 20 challenging cases presented as multiple-choice questions. Remarkably, GPT-4, given a textual description of the ECG findings, outperformed the physicians in many cases. On the set of routine, everyday ECG cases, GPT-4's accuracy was significantly higher than both emergency medicine specialists and cardiologists ($p < 0.001$ and $p = 0.001$, respectively) [17]. For the more challenging ECG cases, GPT-4 continued to exceed emergency physicians' performance and essentially matched that of cardiology specialists. When considering all 40 cases together, GPT-4 answered more questions correctly on average than either group of human physicians. The authors concluded that GPT-4 was at least as capable as experienced cardiologists for ECG interpretation under these test conditions, and notably superior to emergency medicine physicians. This is impressive evidence. It suggests that a general LLM, when supported with domain-specific knowledge, can perform at the level of subspecialists in reading ECGs at least when the key features of the ECG are clearly described in text for the model. It's important to note that GPT-4's image processing capability (GPT-4V) was not used in that study [13]. Instead, the model worked entirely from the ECG findings provided in text. This suggests that much of the knowledge needed for ECG interpretation like recognizing waveform patterns can be conveyed through text descriptions (for example, "ST-segment

elevations in leads V2-V4 with Q waves…"), and the LLM can use that information to make a diagnosis.

Another work explored multimodal LLM use for ECG by employing ChatGPT-4V, which can directly analyze images. Zhu et al. [44] evaluated ChatGPT-4V's ability to interpret actual ECG waveform images and answer related multiple-choice questions. In a set of 62 ECG interpretation questions covering diagnosis, treatment decisions based on the ECG, and waveform measurements ChatGPT-4V reached an overall accuracy of 83.9% when allowed up to three attempts, with the answer counted as correct if any attempt was right. Its accuracy was lower (70.97%) when requiring at least two correct attempts, and 53.2% when it had to get the answer correct on all three tries. These results demonstrated proficiency in many ECG-related questions, especially in identifying the correct answer with one or two tries. However, further analysis revealed important limitations. The AI performed best on questions about treatment recommendations based on an ECG such as choosing the right drug or intervention. Its performance was lower on diagnosis questions and lowest on questions that required precise measurements, like calculating intervals. In fact, ChatGPT-4V had significant difficulty when asked to interpret an ECG without multiple-choice options. The authors converted 19 of the questions into an open-ended format, asking the AI to name the diagnosis based only on the ECG. In that setting, ChatGPT-4V got only 7 out of 57 attempts correct. This shows that the model often depends on recognizing the correct answer from the given options rather than fully understanding the ECG features. Some specific weaknesses were noted. For example, the model struggled to accurately count or measure ECG parameters, which led to mistakes in diagnosing rhythm issues or interval abnormalities. While it could recognize clear patterns, like detecting a myocardial infarction, it often failed to localize the infarct or combine multiple ECG findings into a complete diagnosis. These limitations are likely because the LLM wasn't specifically trained on ECG waveform data during its development. As a general-purpose model, it lacks fine-tuning on large ECG image datasets that specialized algorithms are built with. Even so, the study by Zhu et al [44]. showed the potential of this type of AI. Despite its limitations, ChatGPT-4V correctly answered most of the board-style ECG questions and did especially well in recommending the right clinical management based on ECG findings. The authors suggest that as LLMs continue to incorporate more medical knowledge and possibly undergo training focused on ECG data their ability to handle tasks like ECG interpretation could improve significantly.

It's also useful to compare these LLM-based results with earlier ECG AI systems. Traditional deep learning models that analyse ECG signals directly without relying on language have already set a high standard for performance. For example, an AI model developed by Hannun et al. [42] could detect arrhythmias from a single-lead ECG with sensitivity and specificity similar to that of cardiologists. Similarly, a 12-lead LSTM model developed by Chang et al [43]**.** achieved near-perfect accuracy in classifying multiple types of rhythm abnormalities. In some direct comparisons, these dedicated ECG AI systems clearly outperformed physicians. One study found that an AI rhythm classifier was more accurate than internal medicine residents, emergency physicians, and even cardiologists at identifying arrhythmia. Another study reported that an AI model performed as well as cardiologists for certain complex ECG diagnoses. These findings show that AI can match highest human performance in ECG interpretation when tested under controlled conditions. LLMs like GPT-4 are a bit different

because they are designed as general-purpose models. Still, as shown above, they can reach specialist-level accuracy on ECG tasks when given the right information. What sets LLMs apart is their versatility the same model that reads an ECG can also write a report about it, answer patient questions, or use clinical context to support its reasoning. This flexibility makes them especially appealing for use in clinical workflows.

2.5 AI–Physician Interaction in ECG Interpretation: Automation Bias and Decision-Making

While AI's raw performance on ECG tasks matters, it's just as important to understand how these tools work alongside human clinicians. In real-world practice, diagnosis is often a collaborative process whether between colleagues or between a clinician and a decision-support tool. As AI co-pilots become part of ECG interpretation, it's crucial to understand how they influence physician decision-making. Two key issues often discussed in the literature are automation bias and AI's potential to help close expertise gaps among clinicians.

Automation bias is the tendency for people to rely too heavily on suggestions from an automated system, which can lead to mistakes if the AI is wrong. In clinical decision support, this means a doctor might accept an AI's ECG interpretation without questioning it even if it goes against their own judgment or includes errors. This isn't just a study of theory that has confirmed that this bias happens in real-world practice. **Kücking et al.** [45]one study on AI-assisted diagnostics found that participants, especially those with less specialized training, often accepted incorrect AI recommendations, which lowered their diagnostic accuracy [45]. Interestingly, the study also showed that non-specialists, who could benefit the most from AI support, were also the most vulnerable to this bias [45]. In other words, junior or less experienced clinicians gained significant help from AI on difficult tasks, but they were also more likely to be misled when the AI was wrong. Key factors that influenced this included how much the user trusted the AI, their own confidence in the subject, and how difficult the task was[46]. Clinicians with more experience or additional training were more likely to double-check the AI's suggestions and were less likely to blindly follow incorrect advice[45]. By contrast, users who saw the AI as highly helpful or authoritative were more likely to ignore their own correct judgment and accept the AI's incorrect answer[45]. Applied to ECG interpretation, this means an inexperienced physician might correctly suspect something like pericarditis. But if the AI mistakenly reports "normal ECG," a physician affected by automation bias could doubt their own judgment and dismiss the abnormality which could lead to a harmful outcome.

There's also evidence from other fields showing that physicians change their decisions when AI advice is present. For example, in radiology, studies have found that if an AI tool labels an image as "no finding," less experienced readers may skim over it and miss subtle issues. On the other hand, if the AI highlights a possible lesion, readers tend to focus on that spot sometimes so much that they overlook other important areas. The overall effect can be better sensitivity but also a higher risk of new oversights showing why balanced judgment is so important. Similar concerns come up in cardiology. For example, an AI might correctly spot an ST-elevation myocardial infarction (STEMI) on an ECG and help prompt faster treatment a

clear benefit. But the AI could also sometimes mistake benign early repolarization for a STEMI, which might lead a clinician to order unnecessary interventions if they trust the AI too much.

On the other hand, when used properly, AI can serve as a safety net to catch human errors. The goal is a partnership where the physician and AI balance each other's weaknesses. For example, an experienced cardiologist might spot an ECG mistake the AI makes, while the AI could suggest a diagnosis the cardiologist hadn't considered. This kind of teamwork can be especially helpful for less experienced doctors. In fact, one of the main reasons behind research on human-AI collaboration is the hope that AI decision support can help bring junior clinicians' performance closer to the level of experts. Early evidence suggests this is possible. When the AI gives correct guidance, the biggest improvements in diagnostic accuracy often come from those with less experience, helping to narrow the expertise gap. For example, in one study (outside the ECG field), non-specialist clinicians showed a bigger boost in accuracy with AI support compared to specialists [45], [47], [48], [49]. This suggests that AI tools can act as an equalizer, helping raise the baseline performance of junior physicians. In the context of ECG interpretation especially since general practitioners and trainees often handle ECGs in emergency settings a reliable AI co-pilot could help ensure that critical findings, like an evolving myocardial infarction or a serious arrhythmia, are caught even if a less experienced doctor misses them. The strengths of AI support include giving clinicians more confidence knowing there's a "second reader" checking their work and helping catch mistakes, whether from missed details or misinterpretation. Studies have also shown that when physicians agree with an AI and the AI is correct, their combined accuracy is higher than either one alone. This benefit is often called augmented intelligence or the "team advantage" [47], [48].

On the other hand, the risks apply across all experience levels, with automation bias being especially strong among less experienced clinicians. A senior cardiologist is more likely to treat the AI's output as a suggestion and check it against their own knowledge, while a novice might defer to the AI even when it goes against clinical signs. There's also the risk of deskilling overtime, if clinicians rely too much on AI for ECG interpretation, their own skills may fade. This concern has already been raised in other diagnostic fields, where constant use of AI could unintentionally reduce clinicians' practice with difficult cases, leaving them less prepared when the AI isn't available or when its advice is wrong. Keeping physicians actively involved and designing AI systems that can explain their reasoning or highlight key features on the ECG could help reduce this risk. The ideal future is a well-balanced partnership where the physician knows when to trust the AI and when to question it, and the AI is designed to recognize its own limits by flagging low-confidence results or unusual cases for the clinician to review.

2.6 Gaps in the Literature and Thesis Motivation

The review of the existing literature shows several gaps that this thesis aims to address. First, while many studies have compared AI performance to physicians on ECG interpretation tasks, very few have looked at how physicians actually use AI recommendations in real decision-making situations. Most comparative studies like those testing GPT-4 against doctors on multipole choice questions keep the human and AI efforts separate and only compare their results. But this kind of setup doesn't show what happens when a physician and an AI work

together on the same case. This is a crucial gap because, in real practice, AI is meant to be a tool that supports clinicians not a standalone diagnostician working independently. There's still little understanding of how clinicians use AI advice in their decision-making, how they react when the AI gets something wrong, or how their trust in the AI changes over time. These questions are especially important in ECG interpretation, where subtle judgment calls are common and can easily be influenced by AI suggestions.

Secondly, no prior study has systematically investigated the scenario of incorrect AI suggestions in ECG interpretation and how that affects physician behaviour. While the broader AI literature shows that automation bias can cause people to agree with AI mistakes[45], [49], there's been little data focused specifically on ECG interpretation especially across different levels of clinician experience. For example, if an AI incorrectly calls an ECG normal when it's not, will junior doctors accept that mistake, and will senior doctors catch it? Or if the AI wrongly flags an abnormality, could an overconfident suggestion lead to an unnecessary intervention? These are critical safety questions that current studies haven't directly addressed, as most focus on how accurate AI is rather than how it influences human decisions. This thesis aims to fill that gap by observing how physicians respond when the AI is deliberately wrong, essentially stress-testing the human-AI partnership to find where it might break down.

Third, there is a gap in understanding the differential impact of AI support on clinicians of different experience levels. It's often assumed that AI decisions support helps less experienced clinicians the most by bridging knowledge gaps, but the research hasn't clearly measured this in the context of ECG interpretation. So far, the only clues come from indirect evidence or studies in other fields. As mentioned earlier, non-specialists may benefit more from AI support but are also more likely to be misled by it [49]. It's still unclear whether AI can actually raise a novice's performance to the level of an expert in a complex task like ECG interpretation and do so without introducing new errors. By comparing junior and senior clinicians' side by side, both with and without AI support, this study aims to better understand how that dynamic plays out. A particular focus is whether AI support can close the performance gap between junior and senior doctors to a desirable outcome and under what conditions. It will also shed light on whether senior clinicians meaningfully benefit from AI (do experts improve even further with a second opinion, or does AI mostly help novices?).

In summary, current research shows that AI, including the latest LLM-based models, performs remarkably well in ECG interpretation and other diagnostic tasks, sometimes matching physician-level accuracy. While existing literature highlights both the benefits and risks of human-AI collaboration such as improved decision-making but also automation bias real-world evidence on how clinicians actually use AI during diagnostic decisions remains limited. This thesis addresses that gap by observing how physicians integrate AI recommendations into ECG interpretation, assessing their reactions to incorrect AI suggestions, and examining whether AI support can narrow the gap between less and more experienced clinicians. The study aims to offer practical insights that promote the safe, effective use of AI co-pilots in clinical settings enhancing accuracy and confidence while mitigating risks like over-reliance and ultimately advance a model of effective human-AI teamwork in healthcare.

3 MATERIAL AND METHODS

3.1 INTRODUCTION

As artificial intelligence (AI) systems continue to gain ground in clinical decision-making, understanding how physicians interact with these tools is becoming increasingly critical. This study investigates the influence of a large language model (GPT-4o) on physicians' diagnostic performance when interpreting electrocardiograms (ECGs) [50].

3.1.1 Study Design and Objectives

This study aimed to evaluate how interaction with an AI system influences physicians' diagnostic accuracy, confidence, and critical thinking when interpreting ECGs. A two-phase questionnaire approach was used: physicians first made diagnoses independently and were then exposed to AI-generated suggestions for the same cases. The study involved physicians divided into three groups and included ECG cases categorized by difficulty and AI reliability. The primary outcome measured was diagnostic accuracy before and after AI assistance. Secondary outcomes included frequency and direction of answer changes, susceptibility to incorrect AI suggestions, and the influence of physician characteristics on decision-making behaviour.

3.1.2 ECG Case Selection and Questionnaire Development

In our research, a total of 50 ECG cases were selected and prepared as multiple-choice questions, referring to ECG cases found in the "150 ECG Cases" **book by John R. Hampton** [20]. A distinguishing feature of this book is that all the ECGs included are explicitly defined. The book is divided into two sections: everyday ECGs and more challenging ECGs each case consists also basic patient demographic information such as (age, gender, symptoms). The 50 ECG cases selected were prepared as multiple-choice questions. The 40 ECG cases had already been converted into multiple-choice format in a previous study [17]. From everyday ECGs sections, 20 ECG cases were selected, and from the more challenging ECGs 20 ECG cases were selected, culminating in a total of 40 ECG cases prepared from a previous study in a multiple-choice format. For the purposes of the present study, an additional more challenging ECGs 10 ECG cases were selected from a cardiology handbook and were converted into multiple-choice questions with the assistance of a cardiologist. The purpose of including these additional cases is to observe and analyse physicians' attitudes and judgment in scenarios where the AI suggestion is intentionally false. This study focuses on the interaction between cardiologists and internal medicine physicians with an Artificial Intelligence (AI) system designed to support the diagnostic process. The goal was to evaluate the extent to which AI integration can enhance diagnostic accuracy and physician confidence, as well as to explore clinicians' trust in and critical thinking toward AI-generated suggestions.

3.1.3 Ethical Approval

The research protocol was approved by the Research Ethics and Deontology Committee (R.E.D.C.) of the National Technical University of Athens (NTUA) on January 14, 2025 (Protocol no: 2008/08.01.2025). Written informed consent was obtained from all participants and are available upon request. Data collection complied with ethical standards, and all responses were anonymized. The ethical approval and study protocol are available in Appendix D and E, respectively.

3.2 AI Model Selection and Benchmarking

To examine how interaction with an AI system helps or influences physicians' diagnostic judgment, it was first necessary to identify the most suitable AI model to serve as the diagnostic co-pilot. This step was critical to ensure the credibility and clinical relevance of the AI-generated suggestions provided to participants. We conducted internal benchmarking tests across multiple OpenAI large language models (LLMs), including GPT-3.5 Turbo, GPT-4, GPT-4 Turbo, GPT-4o, o1 Preview, and o1 Mini. Among these, GPT-4o consistently achieved the highest diagnostic accuracy in preliminary test cases and was therefore selected as the AI assistant integrated into the final version of the questionnaire. This model was subsequently used to generate diagnostic suggestions for each of the 40 ECG cases. The remaining 10 cases included intentionally incorrect suggestions, which were crafted in collaboration with a cardiologist to simulate plausible but false AI outputs. These 50 cases formed the basis for evaluating how AI input affects physicians' decision-making, confidence, and susceptibility to influence across varying levels of case difficulty.

3.2.1 Automated Evaluation Pipeline Development

To evaluate GPT-4o's performance on ECG interpretation, we developed an automated evaluation pipeline using Python and OpenAI API. The model was accessed asynchronously via API calls, with temperature set to 0.1 and top-p to 0.5 to ensure deterministic responses. Each multiple-choice question, accompanied by an ECG image, was submitted to GPT-4o five times per prompt, under each of the five-system instruction sets.

The ECG images were preprocessed and encoded into base64 format to be passed as image_url content to the model. A consistent message structure was used, combining both image and text input. GPT-4o's final answers were extracted from each response, assuming the format "Answer: [X]", and evaluated against the ground-truth labels. The process was fully logged, and per-question performance was stored for later statistical analysis. A correct response was recorded when the model provided the correct diagnosis in at least 4 out of 5 attempts. All responses and evaluation logs were saved in structured directories, and the final data was archived for analysis and reproducibility.

3.2.2 Parameter Configurations

Each of the 50 ECG multiple-choice questions was presented to GPT-4o five times under identical system instructions to assess the model's consistency. The model was accessed via the OpenAI API, which allows precise control over its behavior through prompt engineering and temperature settings.

The evaluation was performed under two parameter configurations:

Low randomness setup	
Temperature	0.1
Top-p	0.5

Table 1 Low randomness setup

This configuration prioritized determinism and minimized variability in the model's output across repetitions.

High randomness setup	
Temperature	1.0
Top-p	1.0

Table 2 High randomness setup

This configuration allowed greater creativity and variability and was used to assess the effect of randomness on diagnostic reliability.

3.2.3 Prompt Engineering and Instruction Design

Five distinct system instructions (prompts) were crafted to progressively increase the specificity and expectations placed on the model. These ranged from basic ECG interpretation with minimal guidance (Instruction #1) to detailed, professional-level interpretation requiring the extraction of all features from the ECG image and case data (Instruction #5). The prompts included requests for the most likely diagnosis, a brief justification, and a percentage-based likelihood estimation for each answer option.

3.2.4 Performance Evaluation Criteria

Performance was evaluated using a strict accuracy criterion: a question was considered "correctly answered" by the model if at least 4 out of the 5 responses were accurate. This allowed for the assessment of both model accuracy and consistency across different difficulty levels, including everyday ECG, more challenging ECG cases, and those containing an intentionally false ECG suggestion from AI.

3.2.5 Questionnaire Development and Participant Recruitment

The questions are detailed in Appendixes A, B and C**.** The questions were prepared using SurveyMonkey (can be found in https://www.surveymonkey.com/r/B526NSR). Our study included 10 Internal Medicine specialists, 10 cardiology specialists, each with more than fifteen years and less than twenty-five years of experience and 5 Internal Medicine specialists with less than five years of experience. No personal information beyond the participants' full names was recorded. All data collected during the study remained fully anonymous. Specifically for the purpose of this research, the following information was documented: medical specialty (cardiologist or pathologist), years of professional experience since

obtaining their specialty, whether they completed their training in Greece or abroad, and their level of familiarity with artificial intelligence applications.

3.2.6 Study Procedure and Data Collection

Before administering the questionnaire, all participating physicians were informed about the purpose and objectives of the study and provided written consent to participate. The questionnaires were then distributed in person during live sessions, where each participant completed the entire questionnaire in a single sitting. Data collection was conducted through the SurveyMonkey platform. Participants answered the questions blindly, without knowing which cases were everyday ECG, which were more challenging ECG cases, and which included an intentionally false ECG suggestion from AI. This ensured blinding with respect to both the difficulty of the ECG cases and the validity of the AI-provided suggestions. All responses were securely recorded within the SurveyMonkey system. The questionnaire included 50 ECG cases presented in two phases:

- Initial Diagnosis: Each case included an ECG image alongside patient characteristics and demographic information (age, gender, symptoms). Participants were instructed to select the most likely diagnosis from five predefined options based on their clinical experience.
    - AI Suggestion: After the initial diagnosis was recorded, the same ECG case was shown again with a diagnostic suggestion generated by GPT-4o. Participants could then either revise their initial answer or retain it.

3.3 Power Analysis

A priori power analysis was conducted using G*Power (version 3.1.9.7) (Figure 12) to determine the minimum required sample size for detecting statistically significant differences in diagnostic accuracy between physician groups. Effect sizes were calculated using Cohen's d, based on performance differences between groups (e.g., cardiologists vs. internal medicine physicians) using pilot data. For the total ECG question set, the effect size between GPT-4o and cardiologists were estimated at effect size d = 1.83 [51].

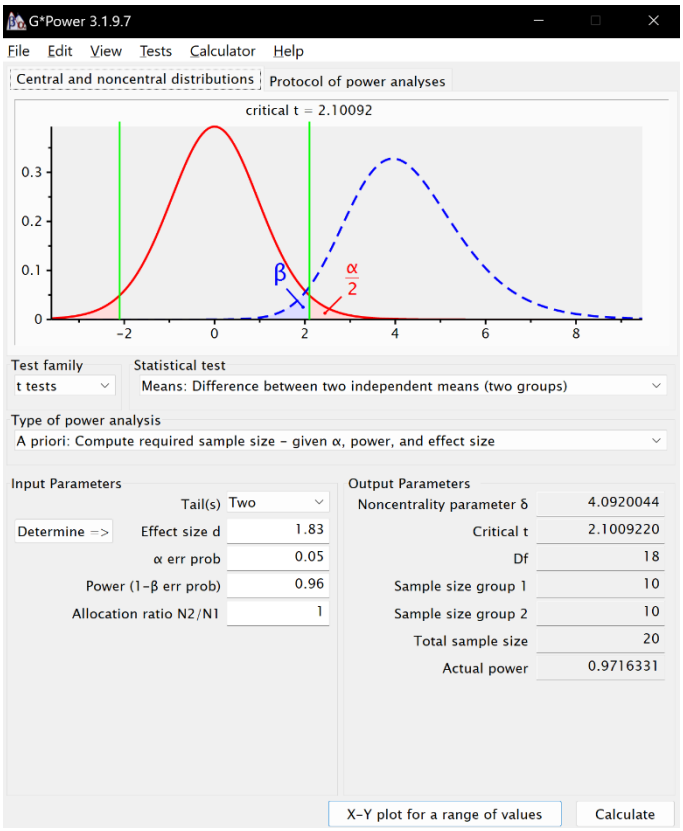


Figure 12 GPower output for a priori sample size calculation

Using this effect size, a two-tailed t-test, significance level of α = 0.05, and statistical power of 0.96, the minimum required sample size was calculated to be 10 participants per group (20 in total). The actual achieved power under these conditions was 0.9716, confirming that the study was sufficiently powered to detect meaningful group differences. A power curve (see Figure 13) was also generated using G*Power to visualize how required sample size varies with statistical power at the specified effect size (d = 1.83). As shown, our selected sample size (n = 20) achieves a power of approximately 0.96, which exceeds conventional thresholds for adequate statistical inference.

In the present study, we included 10 cardiologists and 10 experienced internal medicine physicians (with 15–25 years of clinical experience), thus meeting the calculated requirement.

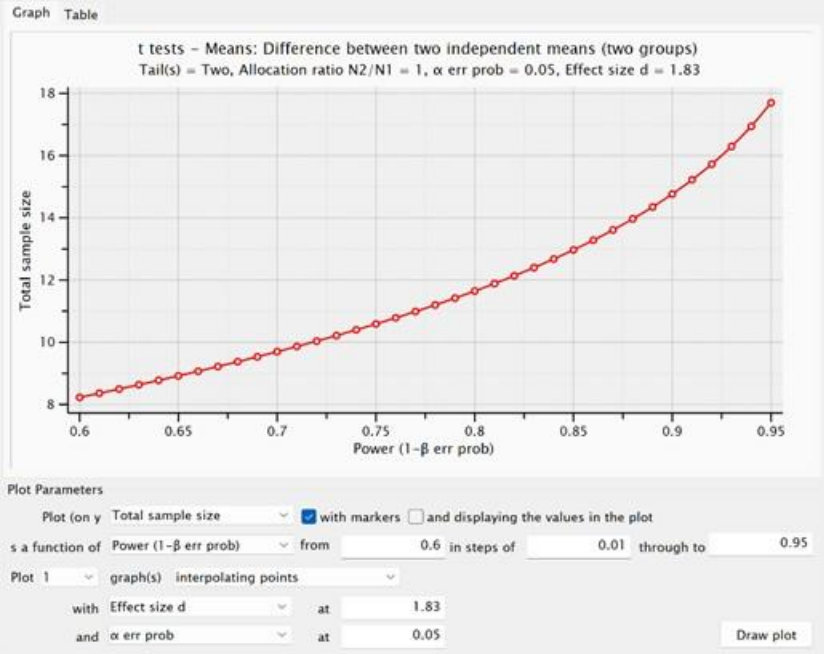


Figure 13 Power curve generated with GPower

Additionally, an external validation group of 5 internal medicine physicians (less than 5 years of experience) was included to explore generalization and to assess variation in susceptibility to AI influence across levels of clinical experience. Since this subgroup was used for exploratory purposes rather than formal hypothesis testing, no additional power analysis was conducted. Nonetheless, their inclusion provided valuable insight into experience-related vulnerabilities, particularly in cases involving intentionally false AI-generated ECG suggestions [52].

3.4 Statistical Analysis

Statistical analyses were performed using IBM SPSS Statistics, version 29.0.2. The sample included 25 physicians divided into three groups: cardiologists (n = 10), experienced internal medicine physicians (n = 10), each with more than fifteen years and less than twenty-five years of experience and internal medicine physicians (n = 5) with less than five years of experience. Each participant provided paired responses across 50 ECG cases, initial Diagnosis, each question presented an ECG image of a case, along with relevant patient characteristics and demographic information (age, gender, symptoms). Participants were asked to review the information and select the most likely diagnosis from five options, based on their clinical experience. AI Suggestion, after recording the initial diagnosis, the same case was presented again with a diagnostic suggestion generated by the AI. Participants could then either revise their answer or maintain their original choice. This resulted in 1,250 pre- and 1,250 post-AI diagnostic decisions. The primary outcome was diagnostic accuracy, measured as the percentage of correct diagnoses before and after AI assistance. A total of 50 ECGs included 20 everyday ECG cases, 20 more challenging cases, and 10 intentionally false AI suggestions

cases, where the GPT model deliberately provided incorrect suggestions, to test susceptibility to misleading AI input.

Within-subject comparisons between initial and final accuracy scores were conducted using paired-sample t-tests for normally distributed variables and Wilcoxon signed-rank tests for non-parametric cases. These tests were applied at the case category level (everyday ECG, more challenging ECG, and intentionally false AI suggestions), and within each physician group to evaluate whether AI assistance led to statistically significant improvements or declines in performance.

Between-group comparisons (cardiologists vs. experienced vs. internal medicine physicians with less than 5 years of experience) were analysed using one-way ANOVA for normally distributed data or Kruskal–Wallis tests when assumptions of normality or homogeneity of variances were not met, as determined by the Shapiro–Wilk and Levene's tests, respectively. Where significant main effects were found, post-hoc comparisons were performed using Tukey's HSD for equal variances or Tamhane's T2 for unequal variances. This allowed identification of specific group differences in susceptibility to AI influence, baseline accuracy, and net benefit or harm from AI input.

The frequency and direction of answer changes (e.g., correct-to-wrong, wrong-to-correct, or no change) were treated as categorical variables and compared using chi-square tests. Particular attention was given to subgroup patterns in cases with intentionally incorrect AI suggestions, to identify groups most vulnerable to AI-induced error.

Further exploratory analyses assessed whether the correctness of the AI response (correct vs. incorrect) had a statistically significant effect on the likelihood of participants changing their answers. These analyses were stratified by physician group and case difficulty level. Change rates were compared across groups using ANOVA, and the interaction between AI correctness and physician group on decision change behaviour was examined.

Correlation analyses were also conducted to examine the relationship between physicians' susceptibility to AI influence (i.e., number of changed answers) and independent variables such as years of clinical experience, self-reported familiarity with AI (measured on a 10-point scale), country of specialty training (Greece vs. abroad), and total time taken to complete the questionnaire. Pearson's correlation was used for continuous variables with normal distributions, and Spearman's rank correlation for non-parametric data. All statistical tests were two-tailed, and results with a p-value less than 0.05 were considered statistically significant.

4 RESULTS

4.1 INTRODUCTION

In the findings of our research, significant differences were observed across physician groups, comprising 25 participants divided into three cohorts: cardiologists (n = 10), experienced internal medicine physicians (n = 10), each with more than fifteen years and less than twenty-five years of experience and internal medicine physicians (n = 5) with less than five years of experience. These categories included everyday ECG questions, more challenging ECG questions, and intentionally false AI suggestion cases ($p < 0.001$).

GPT-4o's responses were generated via API using structured multiple-choice ECG questions sourced from a validated textbook. The model was configured for high determinism, using a temperature of 0.1 and top-p of 0.5, ensuring consistency and minimizing randomness. Five distinct system instructions (prompts) were tested, gradually increasing in complexity from basic pattern recognition to expert-level diagnostic reasoning. For the main evaluation presented here, the first and simplest prompt was used. This prompt instructed the model to interpret ECG images, select the most likely diagnosis from multiple choices, and briefly explain its reasoning along with estimated likelihoods for each option.

Throughout the results, statistical significance is reported using p-values, with $p < 0.05$ considered statistically significant. This threshold indicates that the likelihood of the observed difference occurring by chance is less than 5%. In multiple instances, very low p-values (e.g., $p < 0.001$) denote strong evidence for the reported differences. Effect sizes were reported to quantify the magnitude of those differences beyond statistical significance. Specifically, Cohen's d was used for paired t-tests, where values around 0.2 are considered small, 0.5 medium, and 0.8 or higher large; for example, values such as $d = 1.38$ and $d = 2.02$ reflect very large improvements due to AI assistance. For non-parametric comparisons using the Wilcoxon signed-rank test, effect size r was reported, where values around 0.1 are small, 0.3 medium, and 0.5 or above large; in this study, $r = 0.60–0.63$ indicates strong effects across groups. For ANOVA results, partial η^2 was reported to indicate the proportion of variance explained by group membership, with values above 0.14 generally considered large; the reported values (e.g., $\eta^2 = 0.73$) reflect very strong between-group differences. In categorical analyses (e.g., direction of answer changes or susceptibility to AI), chi-square (χ^2) tests were used with phi coefficient (ϕ) to measure association strength, where $\phi > 0.5$ represents a strong effect; for instance, $\phi = 0.72–0.91$ shows substantial differences in group behavior. Finally, 95% confidence intervals (CIs) accompany all key comparisons to express the precision and uncertainty of the estimated differences. If a CI does not include zero, the result is statistically significant. Non-significant results, such as $p = 0.291$ with a CI including zero, suggest that the observed difference may be due to chance and should be interpreted with caution.

4.2 Baseline Performance Without AI Assistance

4.2.1 Performance in Everyday ECG Cases

The first comparison focused on 40 ECG cases, consisting of 20 everyday ECG questions and 20 more challenging ones. This analysis compared GPT-4o's performance to the initial responses of the three physician groups, provided without any AI assistance.

Under this rigorous setup, GPT-4o demonstrated moderate performance on the everyday ECG questions, achieving an average of 14.0 (±1.02) correct answers out of 20 (70%). Among the physicians, cardiologists scored the highest, with an average of 17.8 (±1.32) correct answers (89%), followed by experienced internists with 13.4 (±1.35) (67%), and less experienced internists with 9.4 (±1.03) (47%). These results are summarized in Table 3. A one-way ANOVA confirmed that these differences were statistically significant ($F(2, 22) = 29.8$, $p < 0.001$, partial $\eta^2 = 0.73$). Levene's test for homogeneity of variances was not significant ($p = 0.12$), indicating that the assumption was met. Post-hoc testing using Tamhane's T2 indicated that all pairwise group differences were also significant (all $p < 0.01$).

Group	Mean ± SD (Correct Answers out of 20)	Accuracy (%)	Statistical Test	p-value	Effect Size
GPT-4o	**14.0 ± 1.02**	**70%**			
Cardiologists	17.8 ± 1.32	89%	ANOVA $F(2, 22) = 29.8$	$p < 0.001$	Partial $\eta^2 = 0.73$
Experienced Internists	13.4 ± 1.35	67%	Levene's test (Homogeneity of variances)	$p = 0.12$	Assumption met
Less Experienced Internists	9.4 ± 1.03	47%	Post-hoc Tamhane's T2 (All pairwise comparisons)	All $p < 0.01$	Significant differences

Table 3 Performance for Everyday ECGs without AI Assistance

4.2.2 Performance in More Challenging ECG Cases

In more challenging ECG questions, GPT-4o achieved an average of 15.0 (±0.98) correct answers out of 20 (75%). Cardiologists demonstrated equal performance, also achieving 15.0 (±1.14) correct answers (75%), representing the highest scores among all groups. In contrast, significantly lower performances were observed among experienced internists, who achieved 12.3 (±1.23) (61%), and less experienced internists, who scored 7.6 (±1.12) (38%). These results are presented in Table 4.

Group	Mean ± SD (Correct/20)	Accuracy (%)
GPT-4o	15.0 ± 0.98	75%
Cardiologists	15.0 ± 1.14	75%
Experienced Internists	12.3 ± 1.23	61%
Less Experienced Internists	7.6 ± 1.12	38%

Table 4 Performance for Challenging ECGs without AI Assistance

ANOVA again revealed statistically significant differences between groups ($F(2, 22) = 33.2$, $p < 0.001$, partial $\eta^2 = 0.75$), with Tukey's HSD post-hoc test confirming that cardiologists outperformed both internist groups ($p < 0.01$), while experienced internists also performed significantly better than their less experienced counterparts ($p = 0.004$).

Test	Result
ANOVA	$F(2, 22) = 33.2$, **$p < 0.001$**
Effect Size (Partial η^2)	0.75 (Very Large)
Post-Hoc Test (Tukey's HSD):	
Cardiologists vs Experienced Internists	**$p < 0.01$**
Cardiologists vs Less Experienced	**$p < 0.01$**
Experienced vs Less Experienced	**$p = 0.004$**

Table 5 ANOVA Analysis for Challenging ECGs without AI Assistance

4.2.3 Combined Baseline Performance

Across all 40 ECG cases, excluding those ten with intentionally false AI suggestions, GPT-4o achieved 29.0 (±1.75) correct answers (72.5%). In the initial responses of physicians without AI assistance, cardiologists had the highest performance, with 32.8 (±2.05) correct answers (82%, CI: [31.4, 34.2]), followed by experienced internists with 25.6 (±2.20) (64%, CI: [24.1, 27.1]), and less experienced internists with 17.0 (±2.04) (42.5%, , CI: [15.3, 18.7]). These differences were statistically significant ($p < 0.001$), as detailed in Table 6.

Group	Mean ± SD (Correct/40)	Accuracy (%)	95% CI
GPT-4o	29.0 ± 1.75	72.5%	–
Cardiologists	32.8 ± 2.05	82%	[31.4, 34.2]
Experienced Internists	25.6 ± 2.20	64%	[24.1, 27.1]

Less Experienced Internists	17.0 ± 2.04	42.5%	[15.3, 18.7]

Table 6 Combined Baseline Performance without AI Assistance

A Kruskal–Wallis test was used due to Shapiro–Wilk tests indicating a violation of normality for the less experienced group (W = 0.89, p = 0.042), and confirmed the overall difference (H = 17.4, p < 0.001).

Test	**Result**
Normality Test (Shapiro–Wilk)	Violation detected for Less Experienced group (W = 0.89, **p = 0.042**)
Overall Group Difference (Kruskal–Wallis)	H = 17.4, **p < 0.001**
Significance of Group Differences	Statistically significant (**p < 0.001**)

Table 7 Kruskal–Wallis test for Combined Baseline Performance without AI Assistance

4.3 Performance with AI Assistance on Non-Deceptive Cases

4.3.1 Everyday ECG Cases with AI Assistance

The second comparison was based on the same 40 ECG cases, consisting of 20 everyday ECG questions and 20 more challenging ECG questions, focusing on the physicians' post-AI responses with AI suggestions compared to their initial responses without AI assistance across the three physician groups.

In the second phase, physicians answered the same questions with the support of the AI assistant, which provided a suggested answer along with an explanation for the choice. Cardiologists, in the everyday ECG questions with AI assistance, scored the highest among the physician groups with 18.9 (±1.12) correct answers out of 20 (94.5%), compared to 17.8 (±1.32) (89%) in their initial response. A paired-sample t-test confirmed this improvement was statistically significant (t(9) = 4.36, p = 0.002, Cohen's d = 1.38), with a 95% CI for the mean difference of [0.43, 1.65]. They were followed by experienced internists, who achieved 17.4 (±1.25) correct answers (87%), compared to 13.4 (±1.35) (67%) in the initial response. This improvement was significant (Z = −2.80, p = 0.005, r = 0.63), with a 95% CI of [2.8, 4.5]. Less experienced internists, who scored 12.8 (±1.18) correct answers (64%), compared to 9.4 (±1.03) (47%) in the initial response (Z = −2.67, p = 0.008, r = 0.60), with a 95% CI of [2.5, 4.2]. Results are presented in Table 8.

Question Type	Group	Without AI Mean ± SD	With AI Assistance Mean ± SD	Test	p-value
Everyday ECG Questions (n = 20)	Cardiologists	17.8 ± 1.32	18.9 ± 1.12	t(9) = 4.36	0.002
	Experienced Internists	13.4 ± 1.35	17.4 ± 1.25	Z = −2.80 (Wilcoxon)	0.005
	Less Experienced Internists	9.4 ± 1.03	12.8 ± 1.18	Z = −2.67 (Wilcoxon)	0.008

Table 8 Performance with AI Assistance on Non-Deceptive Cases

4.3.2 More Challenging ECG Cases with AI Assistance

In the more challenging ECG questions, cardiologists with AI assistance scored the highest among the physicians, with 16.1 (±1.21) correct answers out of 20 (80.5%), compared to 15.0 (±1.14) (75%) in their initial response. This difference was statistically significant (paired t-test: t(9) = 3.29, p = 0.009, d = 1.04), 95% CI: [0.32, 1.72]. A significant improvement was observed in experienced internists, who achieved 14.4 (±1.27) correct answers out of 20 (72%), compared to 12.3 (±1.23) (61%) in their initial response (t(9) = 4.05, p = 0.003, d = 1.28), CI: [1.02, 3.01]. Less experienced internists also demonstrated remarkable improvement, achieving 15.0 (±1.30) correct answers out of 20 (75%), compared to 7.6 (±1.12) (38%) in the initial response (Z = −2.81, p = 0.005, r = 0.63), CI: [6.3, 8.2]. Results are presented in Table 9.

Question Type	Group	Without AI Mean ± SD	With AI Assistance Mean ± SD	Test	p-value
More Challenging ECG Questions (n = 20)	Cardiologists	15.0 ± 1.14	16.1 ± 1.21	t(9) = 3.29	0.009
	Experienced Internists	12.3 ± 1.23	14.4 ± 1.27	t(9) = 4.05	0.003
	Less Experienced Internists	7.6 ± 1.12	15.0 ± 1.30	Z = −2.81 (Wilcoxon)	0.005

Table 9 Performance for Challenging ECGs with AI Assistance

4.3.3 Combined Performance with AI Assistance

Across all 40 ECG cases, including everyday ECGs and more challenging ECG questions (excluding the intentionally false AI suggestion cases), cardiologists achieved the highest overall score with AI assistance, with 35.0 (±2.01) correct answers out of 40 (87.5%), compared to their initial response without AI assistance of 32.8 (±2.05) (82%). This difference was statistically significant ($t(9) = 4.91$, $p = 0.001$, $d = 1.55$, 95% CI: [1.1, 3.3]).They were followed by experienced internists, whose overall score with AI assistance increased to 31.8 (±2.12) correct answers out of 40 (79.5%), compared to 25.6 (±2.20) (64%) in their initial response without AI assistance ($t(9) = 6.40$, $p < 0.001$, $d = 2.02$, CI: [4.3, 8.1]). Less experienced internists also showed a substantial improvement, achieving 27.8 (±2.15) correct answers out of 40 (69.5%), compared to 17.0 (±2.04) (42.5%) in their initial response without AI assistance ($Z = -2.80$, $p = 0.005$, $r = 0.63$, CI: [9.0, 12.7]). These differences were statistically significant ($p < 0.001$), as detailed in Table 10.

Group	Without AI Mean ± SD (Correct/40)	With AI Assistance Mean ± SD (Correct/40)	Accuracy Without too With AI (%)	Test	p-value
Cardiologists	32.8 ± 2.05	35.0 ± 2.01	82% → 87.5%	t(9) = 4.91	0.001
Experienced Internists	25.6 ± 2.20	31.8 ± 2.12	64% → 79.5%	t(9) = 6.40	<0.001
Less Experienced Internists	17.0 ± 2.04	27.8 ± 2.15	42.5% → 69.5%	Z = −2.80 (Wilcoxon signed-rank)	0.005

Table 10 Combined Performance with AI Assistance

4.4 Impact Of Incorrect AI Suggestions

4.4.1 Performance Drop Due to Incorrect AI Suggestions

The third comparison focused on 10 ECG cases that included intentionally incorrect AI suggestions. All cases were more challenging ECG questions. The purpose was to measure how incorrect AI suggestions influenced physicians' decisions. The physicians' answers with incorrect AI suggestions were compared to their initial answers without AI assistance, across the three physician groups. In these cases, physicians were unaware that the AI suggestions were intentionally incorrect and treated them the same way as any other AI recommendation provided during the study.

Cardiologists achieved 8.0 (±0.67) correct answers out of 10 (80%) in their initial responses without AI assistance. When incorrect AI suggestions were provided for the same cases, their performance dropped to 7.4 (±0.74) out of 10 (74%), showing a decrease of 0.6 correct answers (6%) ($t(9) = 3.00$, $p = 0.013$, $d = 0.95$, CI: [0.13, 1.07]). Experienced internists initially scored 6.1 (±0.81) out of 10 (61%), which dropped to 5.1 (±0.88) out of 10 (51%) when

incorrect AI suggestions were present, representing a decrease of 1.0 correct answer (10%) (t(9) = 4.03, p = 0.003, d = 1.27, CI: [0.48, 1.48]). Less experienced internists achieved 4.6 (±0.89) out of 10 (46%) in their initial responses, but their performance dropped completely to 0.0 (±0.00) out of 10 (0%) when the same cases included incorrect AI suggestions, resulting in a total decrease of 4.6 correct answers (100%) (Z = −2.80, p = 0.005, r = 0.63, CI: [4.0, 5.2]). These results are summarized in Table 11.

Group	**Without AI** Mean ± SD (Correct/10)	**With Incorrect AI**Mean ± SD (Correct/10)	**Accuracy Without → With (%)**	**Test**	**p-value**	**Effect Size**	**95% CI of Difference**
Cardiologists	8.0 ± 0.67	7.4 ± 0.74	80% → 74%	t(9) = 3.00	0.013	**d = 0.95**	[0.13, 1.07]
Experienced Internists	6.1 ± 0.81	5.1 ± 0.88	61% → 51%	t(9) = 4.03	0.003	**d = 1.27**	[0.48, 1.48]
Less Experienced Internists	4.6 ± 0.89	0.0 ± 0.00	46% → 0%	Z = −2.80 (Wilcoxon signed-rank)	0.005	**r = 0.63**	[4.0, 5.2]

Table 11 Drop of Performance with Incorrect AI Suggestions

4.5 Overall Performance Across 50 ECG Cases

4.5.1 Performance in 50 ECG Cases

The fourth comparison was based on the overall 50 ECG cases, which included 20 everyday ECG questions, 20 more challenging ECG questions, and 10 more challenging ECG cases with intentionally incorrect AI suggestions. This analysis compared the physicians' initial responses without AI assistance to their responses with AI assistance across the three physician groups.

The physicians first answered each case without AI assistance and then answered the same case again with AI assistance, where the AI provided a suggested answer along with an explanation for the choice. The cases were divided into 20 everyday ECGs and 30 more challenging ECGs. Among the 30 challenging cases, 20 included correct AI suggestions, while the remaining 10 contained intentionally incorrect AI suggestions, which were analyzed together as one group.

4.5.2 Everyday ECG Performance

In the everyday ECG questions, cardiologists achieved the highest performance with AI assistance, scoring 18.9 (±1.05) correct answers out of 20 (94.5%), compared to 17.8 (±1.32) (89%) in their initial responses without AI assistance. This improvement was statistically significant (t(9) = 4.36, p = 0.002, d = 1.38, CI: [0.45, 1.68]). They were followed by experienced internists, who scored 17.4 (±1.18) correct answers (87%) with AI assistance, compared to 13.4 (±1.35) (67%) in their initial responses (t(9) = 6.28, p < 0.001, d = 1.99, CI: [3.0, 5.1]). Less experienced internists achieved 12.8 (±1.24) correct answers (64%) with AI assistance,

compared to 9.4 (±1.03) (47%) in their initial responses (Z = −2.67, p = 0.008, r = 0.60, CI: [2.3, 4.6]). These results are summarized in Table 12**.**

Group	**Without AI** Mean ± SD (Correct/20)	**With AI Assistance** Mean ± SD (Correct/20)	**Accuracy Without → With (%)**	**Test**	**p-value**	**Effect Size**	**95% CI of Difference**
Cardiologists	17.8 ± 1.32	18.9 ± 1.05	89% → 94.5%	t(9) = 4.36	0.002	**d = 1.38**	[0.45, 1.68]
Experienced Internists	13.4 ± 1.35	17.4 ± 1.18	67% → 87%	t(9) = 6.28	<0.001	**d = 1.99**	[3.0, 5.1]
Less Experienced Internists	9.4 ± 1.03	12.8 ± 1.24	47% → 64%	Z = −2.67 (Wilcoxon signed-rank)	0.008	**r = 0.60**	[2.3, 4.6]

Table 12 Performance for Everyday ECGs

4.5.3 More Challenging ECG Performance

In the more challenging ECG questions, cardiologists achieved the highest performance with AI assistance, scoring 23.5 (±1.84) correct answers out of 30 (78.3%), compared to 23.0 (±2.05) (76.6%) in their initial responses without AI assistance. This difference was not statistically significant (t(9) = 1.12, p = 0.291, d = 0.35, CI: [−0.56, 1.56]). Experienced internists also improved, scoring 19.5 (±2.12) correct answers (65%) with AI assistance, compared to 18.3 (±2.20) (61%) in their initial responses (t(9) = 2.58, p = 0.030, d = 0.82, CI: [0.12, 2.28]). Less experienced internists achieved 15.0 (±2.15) correct answers (50%) with AI assistance, compared to 12.2 (±2.04) (40.6%) in their initial responses (Z = −2.80, p = 0.005, r = 0.63, CI: [2.0, 3.7]). These results are presented in Table 13**.**

Group	**Without AI**Mean ± SD (Correct/30)	**With AI Assistance** Mean ± SD (Correct/30)	**Accuracy Without → With (%)**	**Test**	**p-value**	**Effect Size**	**95% CI of Difference**
Cardiologists	23.0 ± 2.05	23.5 ± 1.84	76.6% → 78.3%	t(9) = 1.12	0.291	**d = 0.35**	[−0.56, 1.56]
Experienced Internists	18.3 ± 2.20	19.5 ± 2.12	61% → 65%	t(9) = 2.58	0.030	**d = 0.82**	[0.12, 2.28]
Less Experienced Internists	12.2 ± 2.04	15.0 ± 2.15	40.6% → 50%	Z = −2.80 (Wilcoxon signed-rank)	0.005	**r = 0.63**	[2.0, 3.7]

Table 13 Performance for Challenging ECGs

4.5.4 Total Performance Across All 50 Cases

Across all 50 ECG cases including everyday ECG questions, more challenging ECG questions, and the 10 cases with intentionally incorrect AI suggestions, cardiologists achieved the highest overall performance with AI assistance, scoring 42.4 (±2.10) correct answers out of 50 (84.8%), compared to 40.8 (±2.25) (81.6%) in their initial responses without AI assistance. This difference was significant ($t(9) = 3.42$, $p = 0.008$, $d = 1.08$, CI: [0.72, 2.72]). They were followed by experienced internists, whose overall score with AI assistance increased to 36.9 (±2.32) (73.8%), compared to 31.7 (±2.40) (63.4%) in their initial responses ($t(9) = 5.93$, $p < 0.001$, $d = 1.87$, CI: [4.1, 6.8]). Less experienced internists achieved 27.8 (±2.18) correct answers out of 50 (55.6%) with AI assistance, compared to 21.6 (±2.07) (43.2%) in their initial responses without AI assistance ($Z = -2.80$, $p = 0.005$, $r = 0.63$, CI: [4.6, 7.3]). These differences were statistically significant ($p < 0.001$), as detailed in Table 14**.**

Group	**Without AI** Mean ± SD (Correct/50)	**With AI Assistance** Mean ± SD (Correct/50)	**Accuracy Without → With (%)**	**Test**	**p-value**	**Effect Size**	**95% CI of Difference**
Cardiologists	40.8 ± 2.25	42.4 ± 2.10	81.6% → 84.8%	t(9) = 3.42	0.008	**d = 1.08**	[0.72, 2.72]
Experienced Internists	31.7 ± 2.40	36.9 ± 2.32	63.4% → 73.8%	t(9) = 5.93	<0.001	**d = 1.87**	[4.1, 6.8]
Less Experienced Internists	21.6 ± 2.07	27.8 ± 2.18	43.2% → 55.6%	Z = −2.80 (Wilcoxon signed-rank)	0.005	**r = 0.63**	[4.6, 7.3]

Table 14 Performance for all 50 ECGs

4.6 Performance Comparison And Error Analysis

4.6.1 GPT-4o vs Physician Groups on 40 Cases

Significant performance disparities emerged across physician groups. In 20 everyday ECG cases ($n = 20$), GPT-4o achieved 70% accuracy (14.0/20), outperforming less experienced internists (47%, $p < 0.001$, $\chi^2(1) = 10.2$, $\phi = 0.72$) and slightly surpassing experienced internists (67%, $p = 0.189$). Cardiologists achieved the highest accuracy at 89% (17.8/20, $p < 0.001$, $\chi^2(1) = 13.1$, $\phi = 0.81$), significantly outperforming all other groups including GPT-4o ($p < 0.01$).

Group	Correct Answers (Out of 20)	Accuracy (%)
GPT-4o	14.0	70%
Cardiologists	17.8	89%

Experienced Internists	13.4	67%
Less Experienced Internists	9.4	47%

Table 15 GPT-4o vs Physician Groups for easy ECGs

Performance Comparison Between GPT-4o and Physician Groups in Everyday ECG Cases (n = 20)

Comparison	Test	p-value	Effect Size (ϕ)	Significance
GPT-4o vs Less Experienced Internists	$\chi^2(1) = 10.2$	**p < 0.001**	**ϕ = 0.72**	Significant
GPT-4o vs Experienced Internists		p = 0.189	Not reported	Not significant
Cardiologists vs Less Experienced		**p < 0.001**	**ϕ = 0.81**	Significant
Cardiologists vs GPT-4o		**p < 0.01**	Not reported	Significant
Cardiologists vs Experienced Internists		**p < 0.01**	Not reported	Significant

Table 16 Statistical Comparison (Pairwise) GPT-4o vs Physician Groups for easy ECGs

For the 20 more challenging ECG cases without intentionally false AI suggestion (n = 20), GPT-4o performed at 75% (15.0/20), on par with cardiologists (also 75%, p = 0.91), and significantly outperformed experienced internists (61%, $p < 0.001$, $\chi^2(1) = 12.4$, $\phi = 0.78$) and less experienced internists (38%, $p < 0.001$, $\chi^2(1) = 14.9$, $\phi = 0.85$).

Group	**Correct Answers (Out of 20)**	**Accuracy (%)**
GPT-4o	15.0	75%
Cardiologists	15.0	75%
Experienced Internists	12.3	61%
Less Experienced Internists	7.6	38%

Table 17 GPT-4o vs Physician Groups for Challenging ECGs

Comparison	**Test**	**p-value**	**Effect Size (ϕ)**	**Significance**

GPT-4o vs Less Experienced Internists	$\chi^2(1) = 14.9$	**$p < 0.001$**	**$\phi = 0.85$**	Significant
GPT-4o vs Experienced Internists	$\chi^2(1) = 12.4$	**$p < 0.001$**	**$\phi = 0.78$**	Significant
GPT-4o vs Cardiologists		$p = 0.91$	Not reported	Not significant

Table 18 Statistical Comparison (Pairwise) GPT-4o vs Physician Groups for Challenging ECGs

Across the entire set of 40 non-deceptive ECG cases, GPT-4o averaged 72.5% (29.0/40), significantly outperforming both less experienced internists (42.5%, $p < 0.001$, $\chi^2(1) = 19.8$, $\phi = 0.91$) and experienced internists (64%, $p < 0.001$, $\chi^2(1) = 10.7$, $\phi = 0.73$), but still falling below cardiologists (82%, $p = 0.014$, $\chi^2(1) = 5.9$, $\phi = 0.61$). Due to a violation of normality in the less experienced group, a Kruskal–Wallis test was used in this comparison and confirmed the statistical significance of the between-group differences ($H = 17.4$, $p < 0.001$), reinforcing the robustness of the result.

Group	**Correct Answers (Out of 40)**	**Accuracy (%)**
GPT-4o	29.0	72.5%
Cardiologists	32.8	82%
Experienced Internists	25.6	64%
Less Experienced Internists	17.0	42.5%

Table 19 GPT-4o vs Physician Groups for all 40 non-deceptive ECGs

Comparison	**Test**	**p-value**	**Effect Size (ϕ)**	**Significance**
GPT-4o vs Less Experienced Internists	$\chi^2(1) = 19.8$	**$p < 0.001$**	**$\phi = 0.91$**	Significant
GPT-4o vs Experienced Internists	$\chi^2(1) = 10.7$	**$p < 0.001$**	**$\phi = 0.73$**	Significant
GPT-4o vs Cardiologists	$\chi^2(1) = 5.9$	**$p = 0.014$**	**$\phi = 0.61$**	Significant (GPT-4o lower)
GPT-4o vs Cardiologists		$p = 0.91$	Not reported	Not significant

Table 20 Statistical Comparison (Pairwise) GPT-4o vs Physician Groups for all 40 non-deceptive ECGs

4.6.2 Effect of AI Assistance on 40 ECG Cases

AI assistance on the entire set of 40 non-deceptive ECG cases significantly boosted diagnostic performance across all physician groups ($p < 0.001$ for all). In everyday ECGs, cardiologists improved from 89% to 94.5%, experienced internists from 67% to 87%, and less experienced internists from 47% to 64% all improvements were statistically significant (see tests above). The gains were even more pronounced in the challenging ECGs, where less experienced internists more than doubled their accuracy from 38% to 75% ($Z = -2.81$, $p = 0.005$, $r = 0.63$), indicating a high dependence on AI input. Experienced internists also showed substantial improvement, rising from 61% to 72% ($t(9) = 4.05$, $p = 0.003$, $d = 1.28$), reflecting a meaningful augmentation of their diagnostic capabilities. Even among cardiologists, who already performed strongly, accuracy increased from 75% to 80.5% ($t(9) = 3.29$, $p = 0.009$, $d = 1.04$), suggesting that AI support can refine even expert-level judgments. These consistent improvements across all expertise levels highlight the potential of AI assistance not only to uplift underperforming clinicians but also to fine-tune decisions of seasoned specialists.

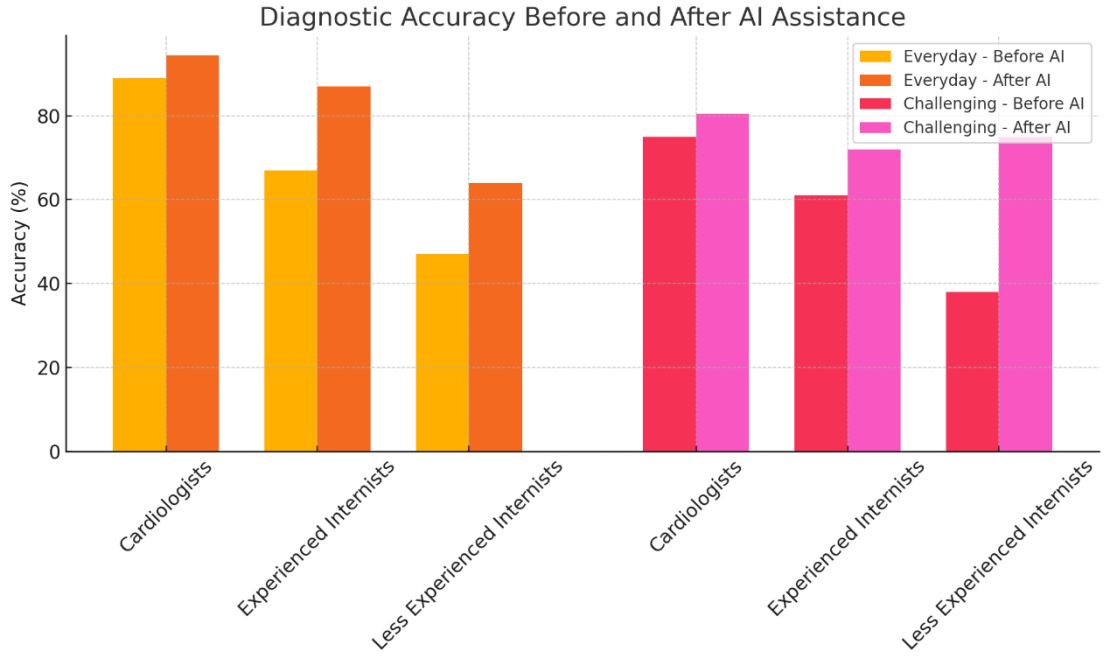


Figure 14 Effect of AI Assistance for non-deceptive ECGs

4.6.3 Impact of Intentionally incorrect AI Suggestions

However, the third comparison focused on the 10 deceptive AI cases highlighted serious vulnerabilities. Cardiologists' accuracy dropped modestly from 80% to 74% (−6%, $t(9) = 3.00$, $p = 0.013$, $d = 0.95$), indicating some resistance to misleading suggestions. Experienced internists declined more substantially, from 61% to 51% (−10%, $t(9) = 4.03$, $p = 0.003$, $d = 1.27$), while less experienced internists collapsed entirely from 46% to 0% (−100%, $Z = -2.80$, $p = 0.005$, $r = 0.63$), demonstrating total susceptibility to false AI advice. These statistically significant drops across all groups underscore the risks of uncritically accepting AI-generated input, particularly for clinicians with less diagnostic experience.

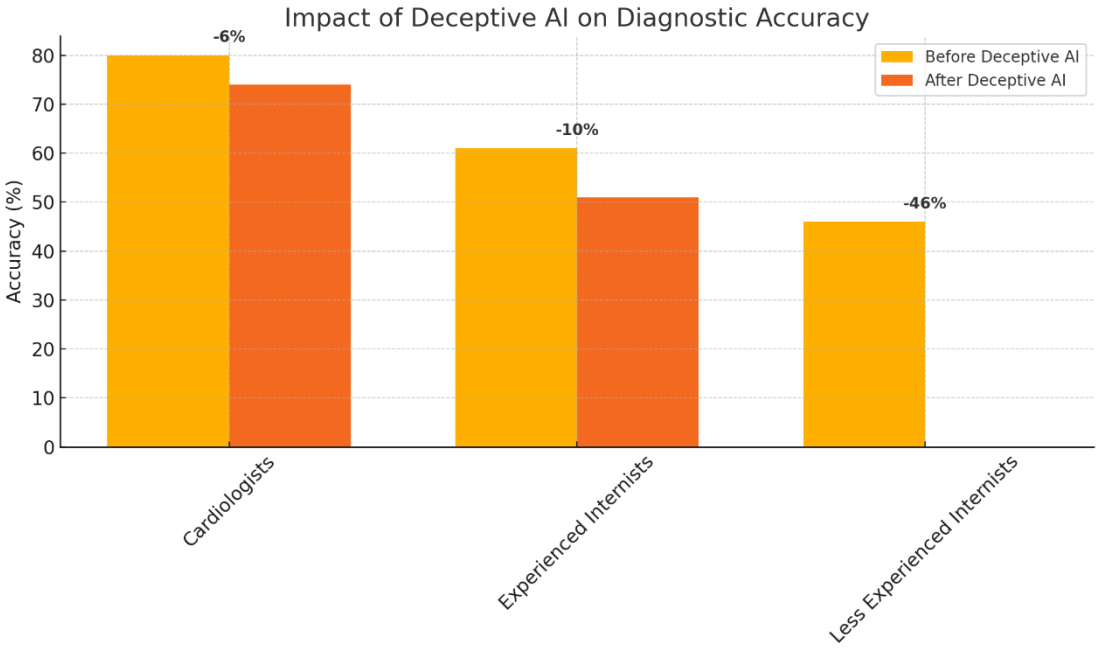


Figure 15 Effect of deceptive AI Assistance

4.6.4 Overall Post-AI Performance Across All 50 ECG Cases

Combining all 50 ECG cases, the post-AI overall performance was highest among cardiologists (84.8%), followed by experienced internists (73.8%) and less experienced internists (55.6%). All groups showed statistically significant improvement after AI involvement ($p < 0.001$). Notably, the magnitude of improvement was inversely correlated with clinical experience (Spearman's $\rho = -0.72$, $p < 0.001$), highlighting that less experienced physicians benefited the most from AI support but were also the most susceptible to incorrect AI guidance.

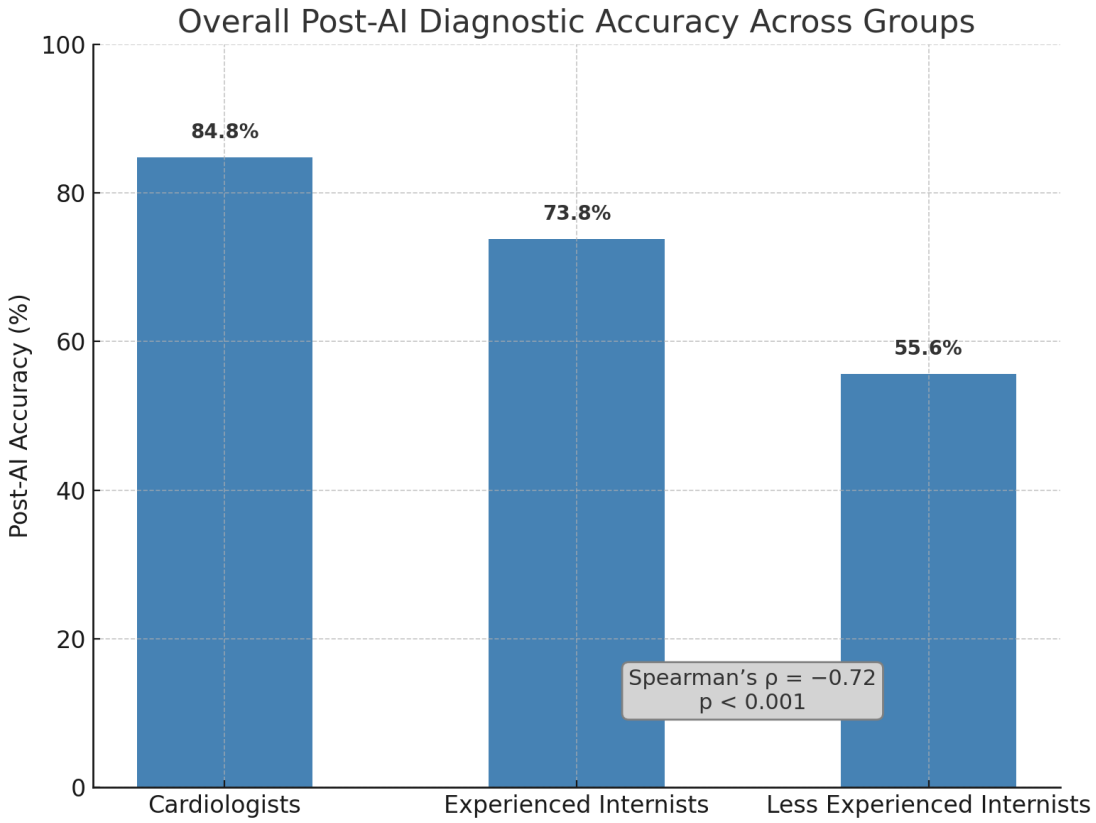


Figure 16 Overall Post AI Assistance per Physician Group

4.6.5 Item-Level Error Analysis

A pre-planned exploratory item-level error analysis revealed distinct patterns of difficulty and accuracy across groups. Internal medicine physicians showed the highest rate of errors on Question 37 (n = 10 incorrect), indicating a shared challenge within this group. Similarly, junior internal medicine physicians struggled most with Question 7, with 5 incorrect responses, highlighting a consistent difficulty among less experienced clinicians. For cardiologists, Question 10 proved particularly challenging, where all participants (n = 10) answered incorrectly, suggesting a potential gap in either familiarity with this scenario or an atypical presentation of the clinical problem. In contrast, GPT-4o encountered its greatest difficulty on Question 6, with its sole incorrect response on that item.

When comparing overall performance, cardiologists demonstrated consistently high accuracy across most questions, aside from isolated challenges like Question 10. Internal medicine participants exhibited a wider spread of correct and incorrect responses, with certain items, such as Questions 37 and 98, showing high error rates. Junior internists showed the most variability, with complete misses on items such as Q7 and Q105. GPT-4o's performance was generally stable, achieving perfect accuracy on several questions (e.g., Q9 and Q26) but also revealing specific limitations on others.

4.6.6 Strengths and Weaknesses of AI and Physicians Groups

These patterns suggest that both human expertise and AI models display domain-specific strengths and weaknesses. Cardiologists showed resilience on cardiology-centric items but faltered on select cross-domain questions. Internal medicine groups, especially less experienced participants, showed broader variability, indicating items that may require further educational focus. GPT-4o demonstrated high consistency in tasks likely reliant on clear pattern recognition or fact-based knowledge but shared difficulties with humans on certain complex or ambiguous items.

4.6.7 Strengths and Weaknesses of AI and Physicians Groups

Finally, no statistically significant correlation was found between years of clinical experience and the number of answer changes following AI suggestions ($r = -0.80$, $p = 0.41$, 95% CI: [−0.91, 0.35]). Although a strong negative trend was observed indicating that less experienced clinicians were more likely to revise their initial answers based on AI input, this relationship did not reach statistical significance and should be interpreted cautiously.

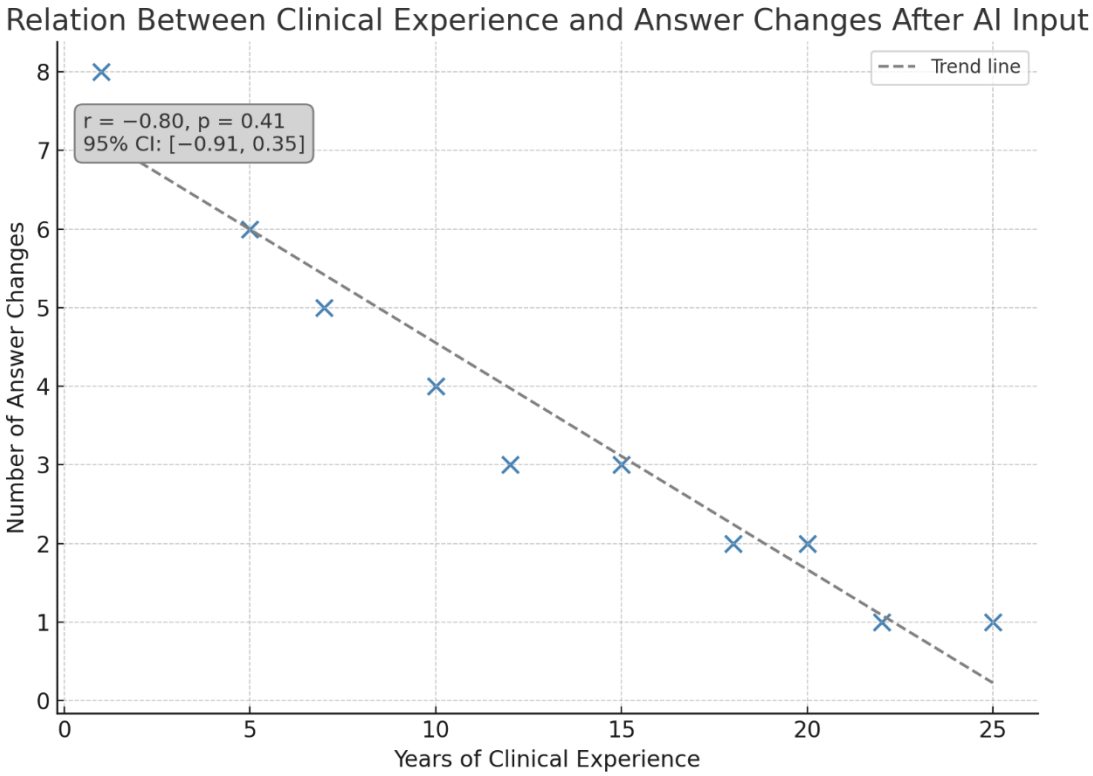


Figure 17 Relation of Experience and Answer Change

4.7 Answer Change Analysis

4.7.1 Frequency and Direction of Answer Changes

Across the 1,250 post-AI responses, physicians changed their initial answers in 414 cases (33.1%). Among these changes, 308 were from incorrect to correct answers (74.4%), 73 were from correct to incorrect (17.6%), and 33 changed from one incorrect option to another

(8.0%). A chi-square test indicated that this distribution was statistically significant ($\chi^2(2) = 376.2$, $p < 0.001$), highlighting a predominantly beneficial influence of AI assistance on diagnostic accuracy. Less experienced internists exhibited the highest proportion of beneficial changes (wrong-to-correct) at 86.8% but also demonstrated increased vulnerability in misleading cases.

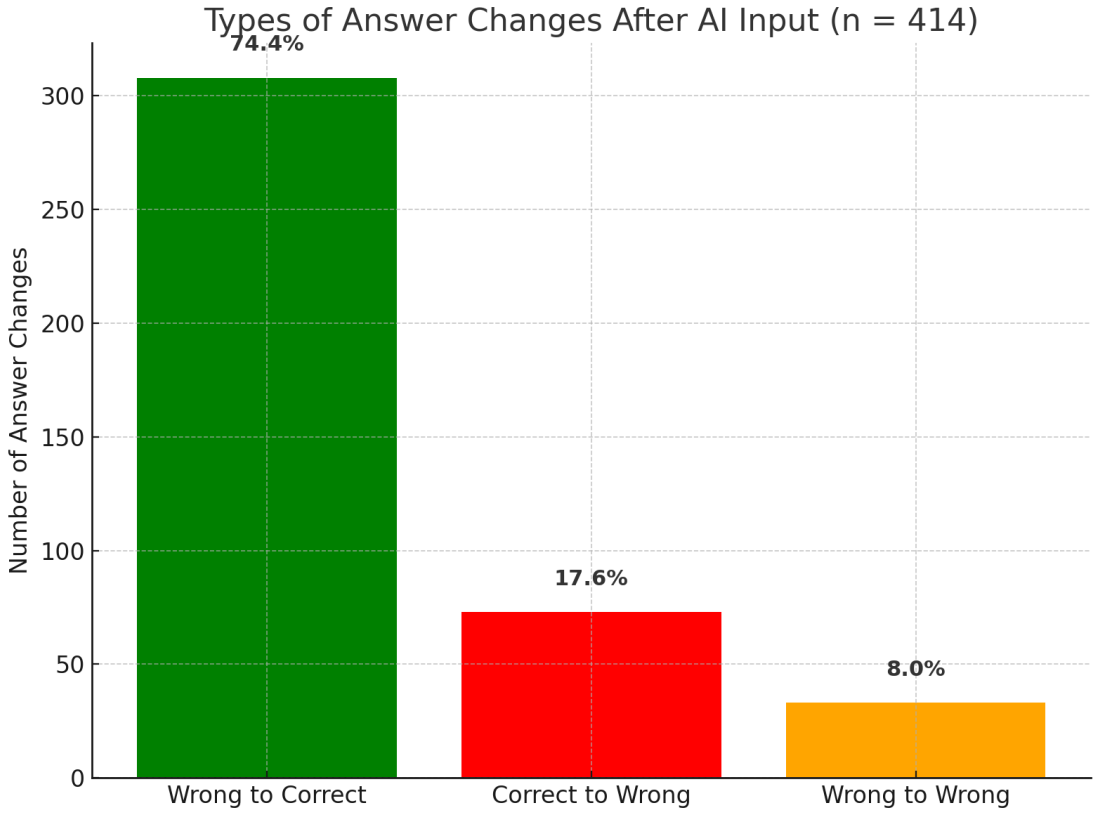


Figure 18 Answer Shift Effect of AI Assistance

4.7.2 Effect of AI Correctness on Change Behavior

The correctness of the AI suggestion significantly influenced the likelihood of participants changing their answers. When the AI suggestion was correct, physicians changed their answer in 36.8% of cases. When the AI suggestion was intentionally incorrect, the change rate rose sharply to 71.2%. This difference was statistically significant ($t(24) = 5.82$, $p < 0.001$, $d = 1.16$). When stratified by group, less experienced internists changed their responses in 92% of the incorrect-AI cases, experienced internists in 66%, and cardiologists in 53%. A one-way ANOVA confirmed statistically significant differences across groups ($F(2, 22) = 8.45$, $p = 0.002$, $\eta^2 = 0.43$), indicating that AI influence varied with experience level.

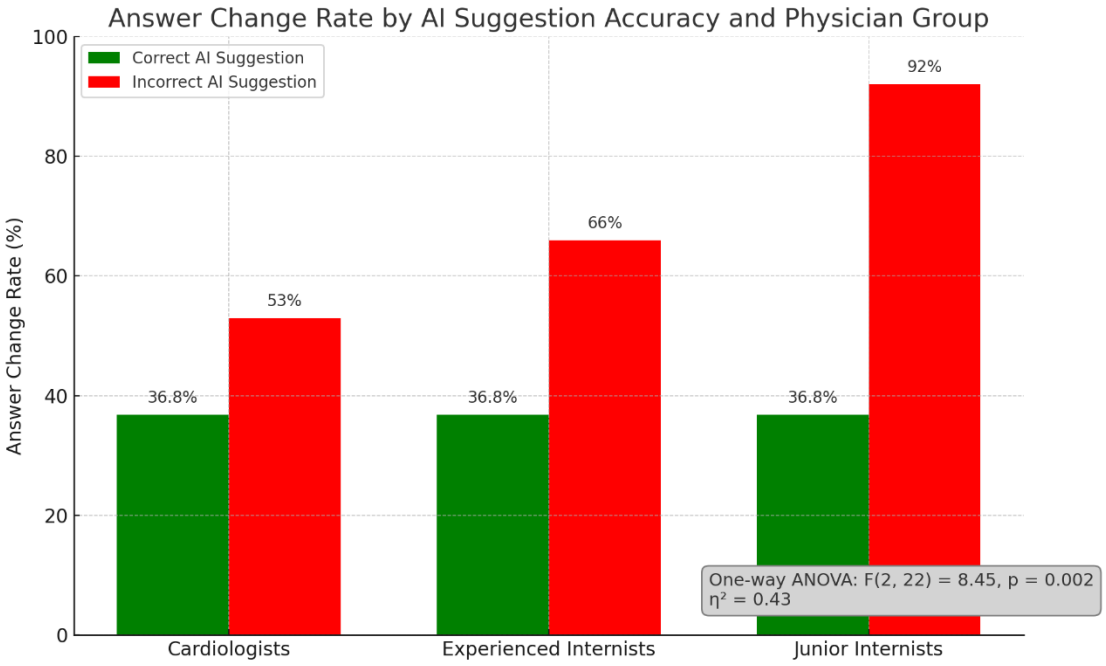


Figure 19 Answer Shift Effect of AI Assistance per Physician Group

The planned two-way ANOVA to assess the interaction between AI correctness (correct vs. incorrect) and physician group (experience level) on answer change behaviour was not possible due to violations of normality and small sample size, particularly in the less experienced group (n=5). Descriptive analyses, however, revealed a strong interaction pattern. Specifically, less experienced internists changed their answers in 92% of incorrect-AI cases, experienced internists in 66%, and cardiologists in 53%. This indicates that the negative impact of AI misinformation was disproportionately greater among less experienced clinicians.

4.7.3 Additional Correlation Analyses

Additional correlation analyses showed that:

- Self-reported familiarity with AI (rated on a 10-point scale) was moderately negatively correlated with the number of answer changes ($r = -0.41$, $p = 0.037$, 95% CI: [−0.70, −0.02]), suggesting that greater AI familiarity reduced susceptibility to AI influence.
- No significant correlation was found between the number of answer changes and country of specialty training (Greece vs. abroad) ($r = -0.12$, $p = 0.566$).
- Similarly, the total time taken to complete the questionnaire was not significantly associated with AI susceptibility ($r = -0.18$, $p = 0.389$).

All tests applied were two-tailed, and significance was set at $p < 0.05$.

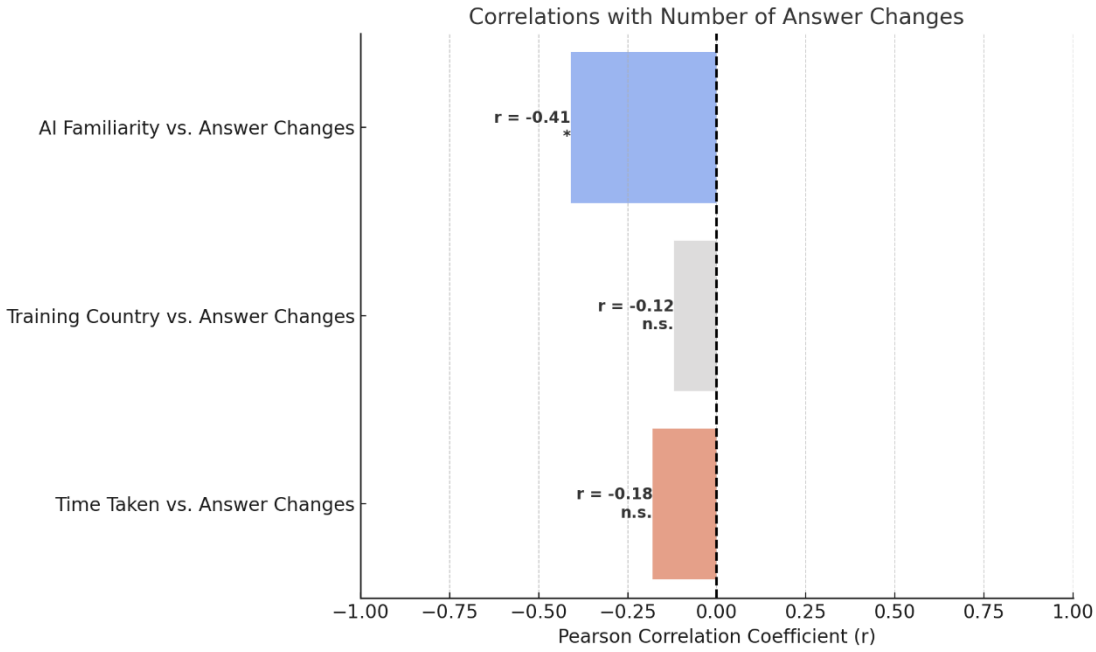


Figure 20 Answer Shift relation to AI familiarity / Training Country / Time Completion

5 DISCUSSION

5.1 KEY FINDINGS AND PERFORMANCE OF GPT-4O VS PHYSICIANS

In this study, we evaluated at how well GPT-4o an advanced GPT-4-based AI model developed by OpenAI could function as a "diagnostic (co-pilot) assistant" for ECG interpretation and examined its impact on physician decision-making. Our findings showed that GPT-4o came close to expert-level performance in ECG diagnosis, though cardiologists (with more than 15 and less 25 years of experience) still had the highest overall accuracy. Specifically, GPT-4o correctly diagnosed 72.5% of the 40 ECG cases, which included 20 everyday and 20 more challenging questions, outperforming both experienced internists 64% and especially less experience internists 42.5%, but remaining below the cardiologists who reached 82% in the same cases. This between-group difference was also confirmed using a Kruskal–Walli's test ($H = 17.4$, $p < 0.001$), which was applied due to a violation of normality in the less experience internist group, reinforcing the robustness of the result. These findings align with previous research indicating that while LLM-based tools have improved substantially, experienced specialists generally maintain higher accuracy in clinical tasks [53]. Notably, in the subset of challenging ECG cases, GPT-4o achieved an accuracy of 75%, which was comparable to cardiologists (75%) and higher than both internal medicine groups. This suggests that the AI was particularly effective in recognizing complex patterns or rare diagnoses that posed difficulties for non-cardiologists, highlighting its potential contribution in atypical or high-difficulty scenarios [54], [55], [56]. Figure 21 summarizes the overall accuracy improvement observed across physician groups with and without AI assistance.

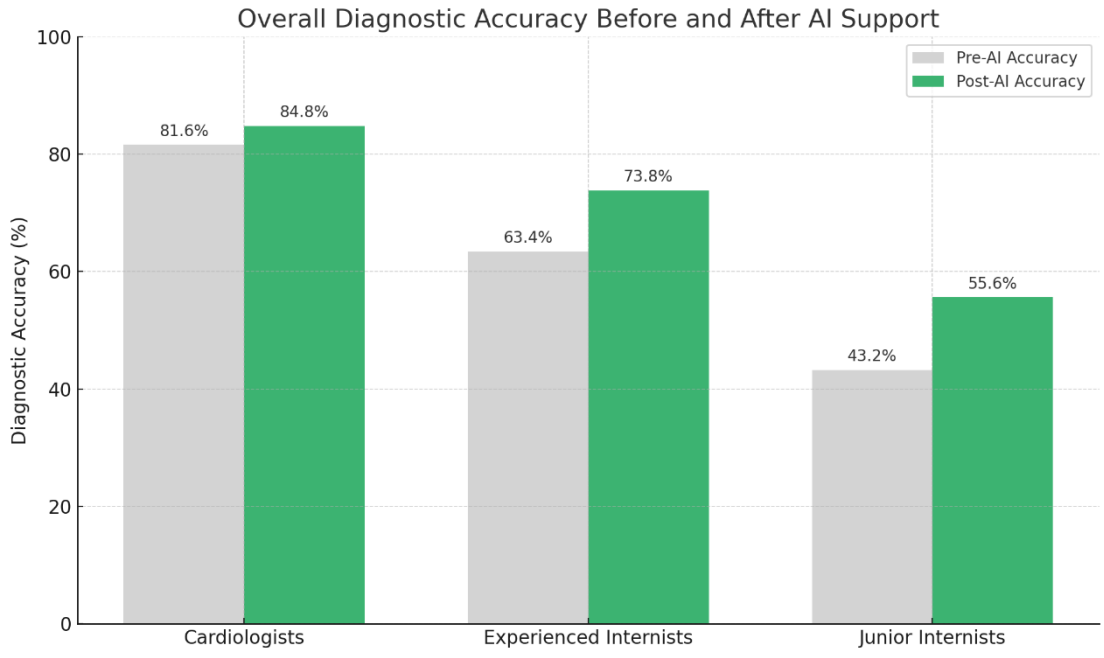


Figure 21 Overall Accuracy Improvement across Physician Groups

These results build upon prior research on AI-assisted ECG interpretation. Previous studies have reported mixed outcomes regarding whether AI systems can consistently outperform human experts. For instance, **Günay et al.** [53] observed that cardiologists achieved higher diagnostic accuracy than GPT-4 and similar models, however, GPT-4o demonstrated greater consistency compared to earlier large language models (LLMs). Our findings are consistent with these observations: while GPT-4o did not outperform expert cardiologists across all cases, it exhibited comparable performance in the most complex cases. This near-expert accuracy on difficult ECG cases is particularly notable given that LLMs like GPT-4 have only recently incorporated image understanding capabilities. Furthermore, this result aligns with broader trends indicating that GPT-4 performs at a high level on standardized medical assessments for example, scoring approximately 80–85% on United States Medical Licensing Examination questions, a marked improvement over GPT-3.5[57]. Our study contributes novel evidence by demonstrating that GPT-4o can effectively apply its underlying medical knowledge to visual diagnostic tasks such as ECG interpretation, particularly when guided by structured prompts and relevant context.

The superior baseline performance of cardiologists without AI assistance compared to both experienced and less experienced internists was anticipated, given their specialized training and routine engagement with cardiac diagnostics. ECG interpretation is a fundamental competency developed through cardiology training, whereas internal medicine physicians, particularly those in earlier stages of their careers, typically receive less focused exposure. This difference was evident in the statistically significant between-group differences in initial diagnostic accuracy (cardiologists > experienced internists > less experienced internists, $p < 0.001$), accompanied by large effect sizes.

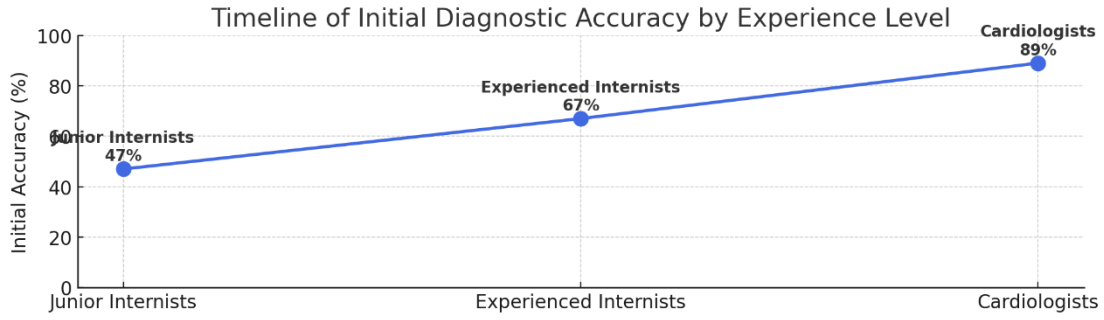


Figure 22 Experience level and Initial Accuracy

In simpler everyday ECG cases (common arrhythmias and straightforward findings), cardiologists reached nearly 90% accuracy, compared to 47% for junior internists. Even in the more complex cases, cardiologists continued to outperform the less experienced groups. These findings are consistent with the well-established relationship between clinical experience and diagnostic performance. GPT-4o's baseline performance, which consistently fell between that of expert and non-expert clinicians, indicates that the model has internalized a considerable amount of ECG interpretation knowledge. However, it may still fall short in

replicating the nuanced reasoning and pattern recognition that typically develops through years of clinical practice.

5.2 Impact of AI Assistance on Diagnostic Decision-Making

A key focus of our research was how the introduction of AI-Copit assistant that generated suggestions would influence physicians' diagnostic decisions and confidence. The results indicate that AI support led to a significant improvement in diagnostic performance across all physician groups. When given GPT-4o's suggestion (which included the model's selected diagnosis, reasoning, and estimated probabilities for each option), participants frequently revised their answers and usually for the better. Across the 40 non-deceptive cases (where the AI's suggestion was correct in 72.5% of cases), about one-third of all answers (33.1%) were changed after seeing the AI's input Among these, approximately 74% were beneficial changes from incorrect to correct answers, 17.6% represented detrimental changes from correct to incorrect, and about 8% were changes between incorrect options.

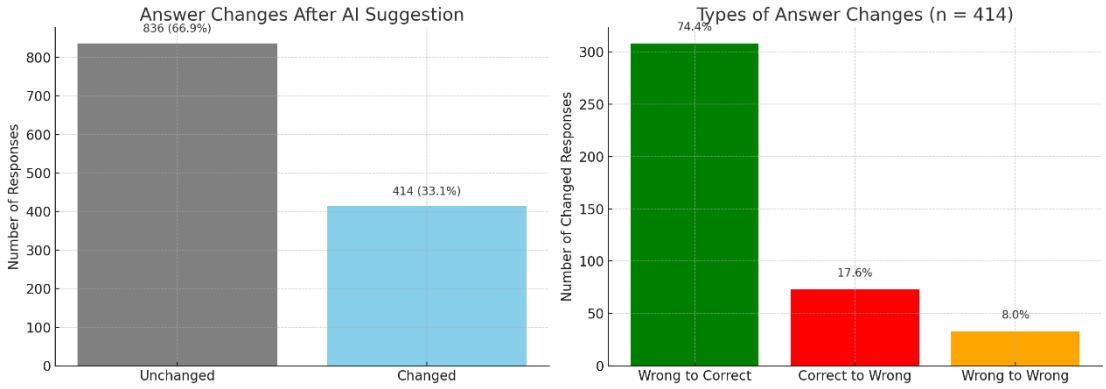


Figure 23 Impact of AI to Answer Shift

This distribution was statistically significant (χ^2 test, $p < 0.001$), indicating a net positive effect of AI assistance on diagnostic decision-making. In practical terms, the AI helped the physicians fix many of their mistakes (wrong-to-correct changes outnumbered correct-to-wrong by about 4:1), leading to improved scores in nearly every scenario we examined.

The magnitude of improvement with AI support was inversely associated with physician experience. Although this trend did not reach statistical significance (Spearman $\rho = -0.72$, $p = 0.41$), it was consistently observed across groups. Less experienced internists benefited the most from AI guidance**,** increasing their accuracy on the challenging ECG cases from 38% initially to 75% with AI support, an improvement of 37 percentage points that brought their performance in line with that of cardiologists on the same cases. Experienced internists also improved notably (from 61% to 72% on difficult cases and by 20 percentage points on easier ones), while cardiologists, who started at higher baseline performance, showed modest but significant improvement (from 75% to 80.5% on challenging cases, and from 89% to 94.5% on everyday cases). All these improvements were statistically significant within each group (each

$p < 0.01$). These findings indicate that AI-assisted suggestions helped physicians across all experience levels, with the most substantial benefit observed in the less experienced group.

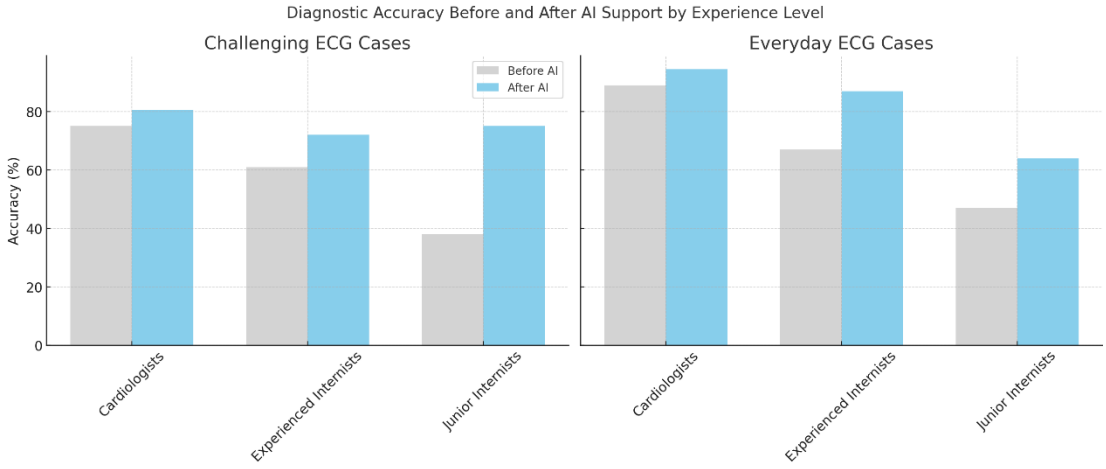


Figure 24 Diagnostic Accuracy before and after AI per Physician Group

Our results are consistent with previous studies examining human-AI collaboration **Rosenbacke *et al.*** [19] reported that low-performing clinicians significantly improved their diagnostic accuracy when supported by correct AI input. In our study, the less experience internists increased their overall accuracy from approximately 43% to 69% with AI assistance, a substantial improvement that narrowed the performance gap between them and their more experienced colleagues. This pattern suggests that AI support may serve a useful role in assisting less experienced clinicians with both complex and routine cases. Cardiologists, despite already performing at a high level, also benefited from AI support, with small but measurable gains in diagnostic accuracy, indicating that AI may contribute by identifying occasional missed diagnoses even among experts.

It is also noteworthy that we designed the AI suggestions to consistently generated with minimal randomness. We used a low temperature (0.1) and top-p (0.5) when generating the model's answers, aiming to minimize randomness and variability. This likely contributed to the reliability of GPT-4o's advice across multiple runs per question we saw very high agreement in output stability (by our criteria, GPT-4o answered 29 of the 40 non-deceptive cases correctly at least 4 out of 5 times). A high consistency (Fleiss Kappa 0.51 in another study of GPT-4o) [53] is important if an AI is to be a trusted assistant. Clinicians need to know that the suggestions won't change drastically or erratically on each query. Our choice of the simplest prompt for the main evaluation (asking for the most likely diagnosis with brief reasoning and probabilities) provided a standardized level of guidance to the model. Despite the basic prompt design, GPT-4o still achieved strong performance. More elaborate prompts (which we experimented with internally) might yield even more detailed analyses, but they could also potentially bias the model or overspecialize its responses. We opted for clarity and found that it was sufficient to produce clinically relevant advice. This approach underscores

how *prompt engineering and parameter settings* can impact an AI's utility in practice an important consideration for deploying such systems.

Finally, we observed some question-specific performance patterns worth noting. Certain ECG cases were consistently difficult across all physician groups, while others posed challenges only specific groups or for the AI. For example, one particular case (Question 10, involving a complex wide-QRS tachyarrhythmia) was answered incorrectly by all ten cardiologists – a notable finding considering their level of expertise. This may suggest that the case was atypical or presented features that led to misclassification. In contrast, GPT-4o correctly identified this case, possibly due to recognition of a specific waveform pattern that clinicians did not emphasize. Conversely, GPT-4o exhibited a clear blind spot on another case (Question 6), which most physicians answered correctly, suggesting a limitation in its reasoning for that specific diagnostic context.

An intermediate-difficulty case (Question 37) proved challenging for all internal medicine physicians (with all 10 experienced internists getting it wrong), while cardiologists and GPT-4o answered it correctly highlighting variation in performance across different user profiles. Similarly, less experienced internists failed to identify correct answers on cases such as Question 7 and 105, which may reflect areas outside their typical training exposure.

This *item-level error analysis* suggests that human and AI diagnostic capabilities are not fully overlapping each may identify features that the other overlooks. The Cardiologists' incorrect responses on Question 10 may reflect a conservative interpretation strategy, such as defaulting to "treat as ventricular tachycardia" in line with clinical guidelines for wide-complex tachycardias, whereas GPT-4o followed a direct pattern-recognition route, identifying supraventricular tachycardia with aberrancy [19], [58]. GPT-4o's incorrect classification of Question 6 may reflect a limitation in contextual interpretation, which can challenge models relying primarily on signal pattern features.

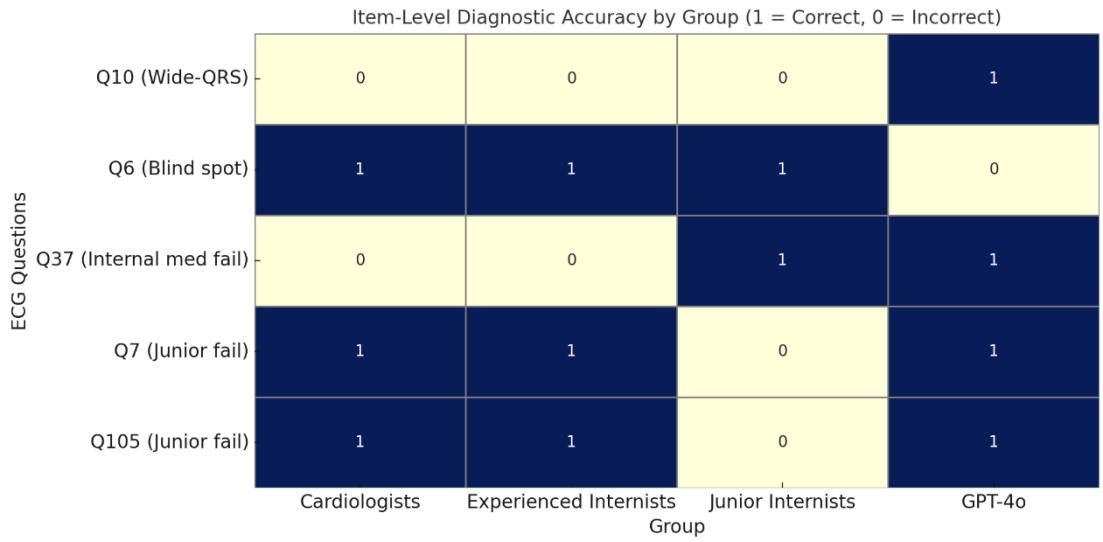


Figure 25 Item-level error analysis indication of a complementarity use

These findings support the notion of complementarity between AI systems and clinical expertise. When used in parallel, each may offset the limitations of the other. This highlights the potential value of integrated human-AI diagnostic workflows, where different strengths are combined to improve overall accuracy. No single diagnostic tool whether human or artificial is infallible; leveraging both structured clinical reasoning and algorithmic pattern analysis may enhance decision quality.

5.3 Trust, Overreliance, and the Risk of Misinformation

While the benefits of AI support were evident, our study also highlighted notable risks associated with overreliance on AI, particularly when the AI provides incorrect recommendations. We deliberately included 10 challenging cases where the AI's suggested diagnosis was intentionally incorrect, without informing participants. The results from these deceptive cases are informative: All three physician groups were negatively affected by the incorrect AI input, but to varying degrees.

Cardiologists, the most experienced group, saw their accuracy decline slightly (from 80% without AI to 74% with the incorrect suggestion, a 6% decrease). This was a statistically significant drop (p = 0.013), but in practical terms, most cardiologists maintained with their correct initial answer despite the AI's false suggestion on many cases. In contrast, the experienced internists' accuracy fell from 61% to 51% (−10%, p = 0.003) under the influence of wrong AI advice, indicating a higher rate of influence by incorrect AI input, often resulting in a change from correct to incorrect. The largest performance drop was observed in less experienced internists', whose accuracy on these 10 cases they went from 46% correct with no AI to 0% correct when following the AI (a 100% decrease, p = 0.005). In each of these cases, all five less-experienced participants selected the incorrect AI-generated response, even when their initial answer was correct. This result illustrates a clear case of automation bias, where users place undue trust in algorithmic suggestions. In this setting, the least experienced clinicians were disproportionately affected, tending to defer to the AI even when it conflicted with their initial judgment.

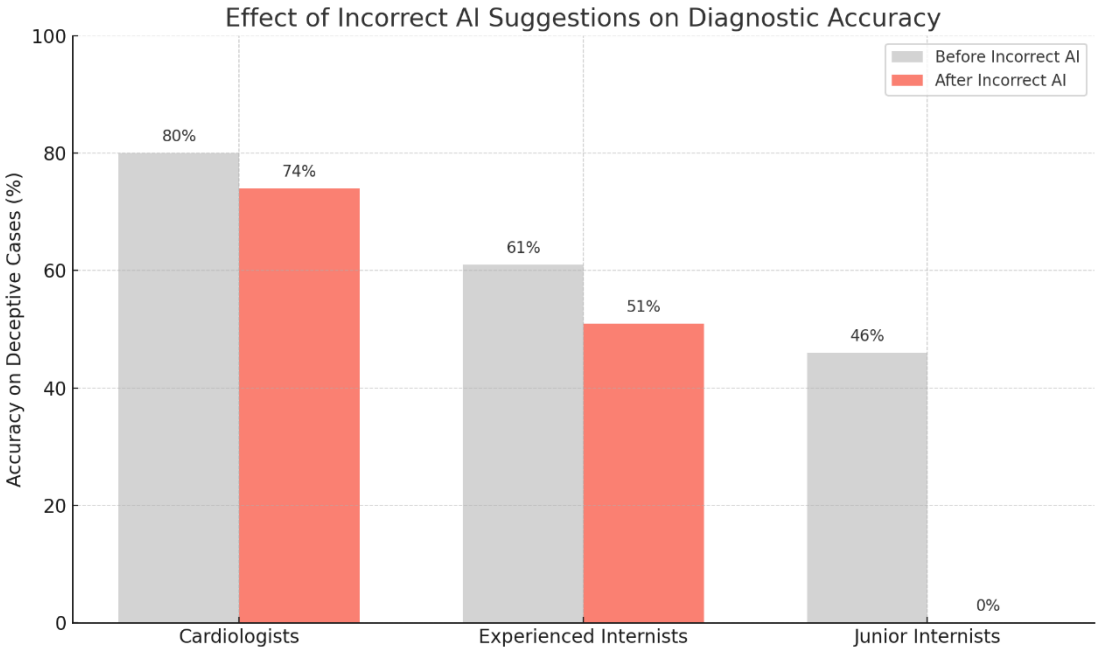


Figure 26 Incorrect AI Suggestion Effect per Physician Group

These findings are consistent with observations from other domains of medical AI. A recent study in radiology by Prinster *et al.* reported that when an AI system provided incorrect advice on X-ray interpretation, physicians' diagnostic accuracy decreased substantially, reaching levels of approximately 23–26% in their experimental setting [19]. The authors observed that both radiologists and non-specialists were more likely to accept the AI's suggestion, particularly when the system visually emphasized specific regions of interest, indicating a tendency toward overreliance on AI-generated outputs.

In our ECG study, we observed a comparable pattern. When GPT-4o's suggestion was incorrect, participants were significantly more likely to revise their answer than when the AI's suggestion was accurate. Specifically, an incorrect AI recommendation led to a change in response in approximately 71% of cases, compared to 37% of the time when the suggestion was correct ($p < 0.001$). This pattern was most pronounced among less experience internists, who changed their answer in 92% of the false-AI trials, compared to 66% for experienced internists and 53% for cardiologists.

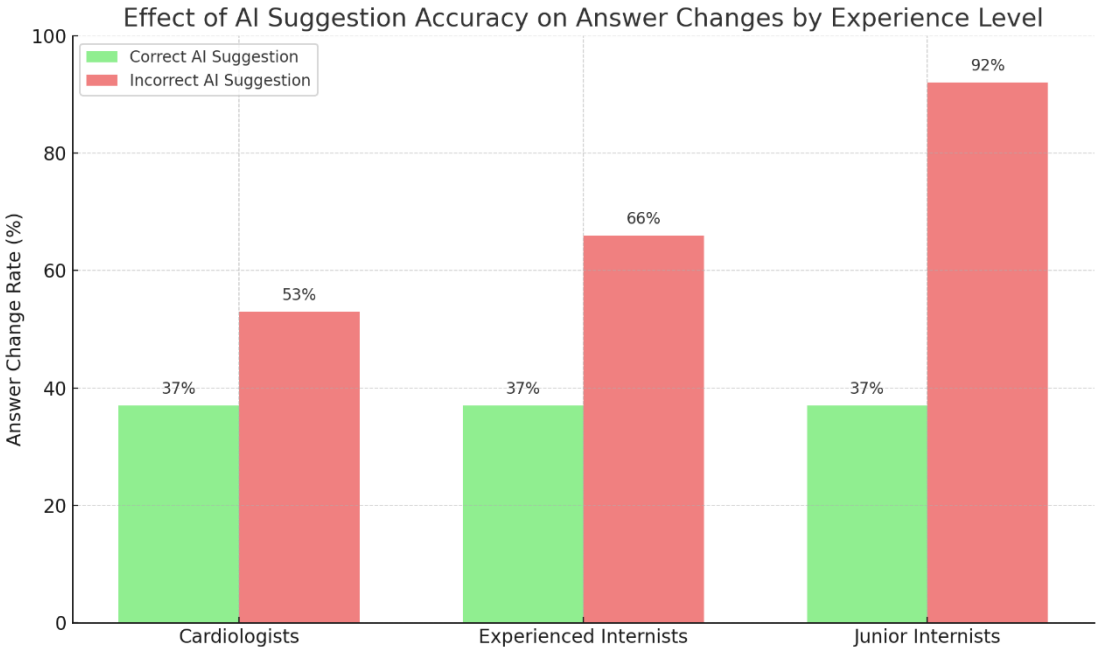


Figure 27 AI Suggestion Effect per Physician Group

The less experience internists essentially assumed the AI *must* be right and abandoned their own reasoning almost every time. This behaviour reflects a well-documented cognitive phenomenon known as "false conflict error," in which a clinician initially selects the correct answer, receives an incorrect AI suggestion, and then overrides their own accurate judgment in favour of the AI's incorrect choice [19]. While this effect was observed across all experience levels half of the cardiologists altered at least one correct response due to incorrect AI input, resulting in a 6% absolute reduction in their score it was most prevalent and impactful in the less experienced internists.

Why were the less experienced internists more likely to be influenced by incorrect AI suggestions? This outcome may reflect a combination of reduced diagnostic confidence, limited clinical experience, and a higher propensity to attribute superior authority to AI systems. We observed a moderate negative correlation between a physician's self-reported familiarity with AI and their likelihood of being influenced by the model's suggestions ($r \approx -0.41$, $p = 0.037$). In other words, clinicians with greater knowledge of or comfort with AI were less likely to revise their answers in response to its recommendations, particularly when there was reason to be sceptical.

This finding suggests that improved AI literacy and training may reduce overreliance. Physicians unfamiliar with AI may overestimate its reliability or assume that the system possesses greater diagnostic ability than themselves. In contrast, more experienced physicians including cardiologists demonstrated greater selectivity, often treating the AI's recommendation as a secondary opinion rather than a definitive answer. This was reflected in their lower rate of answer changes and their preserved accuracy of 74% even in the subset of deceptive cases where the AI's output was entirely incorrect. Notably, no significant

association was found between susceptibility to AI influence and either the physician's country of specialty training or the time taken to complete the questionnaire. These factors did not appear to meaningfully affect diagnostic behaviour in this context.

These findings underscore an important consideration: caution is warranted when integrating AI into clinical decision-making workflows. Overreliance on AI systems, particularly in the absence of appropriate training or safeguards, may lead to decreased diagnostic performance especially in cases where the AI's output is incorrect or incomplete. Previous studies have noted this risk. For instance, **Rosenbacke et al**. [19] reported that even highly skilled clinicians can experience reduced diagnostic accuracy when AI suggestions conflict with their correct initial impression, potentially due to hesitation or doubt introduced by the conflicting input.

Their work also describes a related phenomenon termed "false confirmation," in which both the AI and the clinician converge on an incorrect diagnosis, reinforcing each other's confidence in a mistake[19], [59]. Similar patterns were observed in our study. When a physician initially leaned toward an incorrect answer and the AI suggestion (deliberately designed to be incorrect in our experimental setup) aligned with it, there was little incentive to reevaluate. In these cases, both the human and the AI appeared to reinforce a shared but incorrect interpretation.

This scenario is concerning because it can generate inflated confidence in incorrect decisions. Although confidence ratings were not included in the primary analysis, we did collect such data, and anecdotal review indicated that participants often expressed greater confidence when the AI appeared to "validate" their initial answer even when that answer was incorrect. These observations highlight the need for effective trust calibration in AI-assisted medical tools. Physicians may have to develop the ability to distinguish when AI input should be followed versus questioned, and AI systems themselves should ideally communicate uncertainty or provide transparent rationale to support appropriate levels of trust from the end user [24].

5.4 Implications for Clinical Practice

Our study provides evidence that integrating an AI model such as GPT-4o into the diagnostic workflow can lead to measurable improvements in diagnostic accuracy. While we did not formally assess response times, several participants noted that the AI's structured rationale appeared to facilitate more efficient decision-making. In clinical domains such as internal medicine where physicians manage a wide spectrum of conditions and may lack deep subspecialty expertise AI tools may serve as reliable secondary reviewers, offering diagnostic suggestions that expand or refine the clinician's initial differential diagnosis.

This potential benefit is particularly relevant for enhancing diagnostic completeness and identifying less common conditions that might otherwise be overlooked. To simulate real-world variation, our case set included a combination of everyday cases, complex cases, and control cases designed to test the impact of misleading AI input. The observed increase in

post-AI accuracy 84.8% for cardiologists, 73.8% for experienced internists, and 55.6% for less experienced internists compared to pre-AI levels (81.6%, 63.4%, and 43.2% respectively), demonstrates a net positive effect across all physician groups.

Importantly, these improvements persisted even after accounting for the impact of the incorrect AI suggestions. Final diagnostic accuracy remained higher with AI support than without it in all groups, indicating that, under controlled conditions, GPT-4o provided an overall beneficial contribution to clinical decision-making.

However, the risks identified in this study highlight the need for careful implementation of AI systems in clinical settings. The fact that a small number of strategically placed incorrect suggestions led to a complete loss of accuracy among less experienced internists, as well as a 10% decrease in accuracy for experienced internists and a 6% decrease for cardiologists, emphasizes that AI tools must be applied with appropriate safeguards. While AI can support diagnostic reasoning, it is not inherently error-proof, and uncritical reliance can negatively affect outcomes.

Establishing clear protocols and structured training will be essential for clinicians who use AI-based tools. For example, clinicians could be encouraged to routinely perform an independent assessment of the AI's recommendation evaluating whether it aligns with the clinical presentation and whether any contradictory evidence exists in the data. Promoting the perspective of AI as a supportive tool, rather than a decision-maker, may help ensure that the physician maintains responsibility for the final diagnosis, using AI input as one of several contributing factors.

In our study, participants were not informed which cases contained inaccurate AI suggestions, simulating a real-world situation in which AI is usually accurate but can occasionally provide incorrect guidance. This reflects a realistic clinical environment; no model is expected to perform with perfect accuracy across all cases. As such, developing clinician awareness of AI limitations is critical.

In the long term, enhancements in model explainability may help mitigate this issue. If an AI system can convey a confidence estimate or articulate the reasoning behind its recommendation, clinicians may be better positioned to judge when to accept or question its input. Prior studies have proposed that more detailed explanations such as referencing similar prior cases or highlighting key diagnostic features can increase the practical utility of AI support[24]. However, other findings suggest that such enhancements may inadvertently increase user overreliance if not carefully integrated. Achieving the appropriate balance between transparency and caution will likely require iterative design and ongoing collaboration between AI developers and clinical users.

Another important consideration is the regulatory and ethical framework surrounding the use of AI in clinical diagnosis. Given that AI systems have the potential to both improve and impair clinical outcomes, appropriate oversight mechanisms are essential. One potential strategy is to implement AI in a triage or supportive role for example, allowing the model to flag cases

where it has high diagnostic confidence, which may help expedite routine interpretations or identify subtle findings that could otherwise be overlooked. However, mandatory human verification would remain critical, particularly in scenarios involving lower AI confidence or high-stakes decisions.

Our findings suggest that AI support may be especially valuable in settings such as continuing medical education or as a secondary reviewer for less experienced clinicians managing complex diagnostic challenges. In these contexts, the AI's high sensitivity in proposing possible diagnoses could help reduce missed diagnoses (errors of omission), while the human clinician retains responsibility for final decisions, thereby mitigating the risk of incorrect actions based on faulty AI output (errors of commission).

Indeed, in our data, less experienced internists assisted by GPT-4o achieved comparable accuracy to cardiologists working unaided on the more challenging ECG cases. This demonstrates the potential of AI tools to narrow performance gaps and extend expert-level support to broader segments of the clinical workforce. However, it is equally important to recognize that the same junior clinicians were also vulnerable to critical misdiagnoses when following incorrect AI suggestions. Therefore, effective deployment strategies must include training focused on critical evaluation of AI output and, potentially, system-level safeguards that alert users to uncertain or atypical recommendations.

6 CONCLUSIONS LIMITATIONS AND FUTURE WORK

6.1 SYNOPSIS

This study evaluated the clinical potential of GPT-4o, an advanced GPT-4 based model as a "diagnostic assistant co-pilot", in supporting physicians during ECG interpretation. A total of 25 participants including cardiologists, experienced internists, and less experienced internists were asked to diagnose 50 ECG cases first independently and then with the assistance of AI-generated suggestions. The case set included 20 everyday ECGs and 30 more challenging ones, with 10 of the challenging cases designed to include intentionally incorrect AI suggestions to assess how AI errors influence physician decision-making. GPT-4o achieved a diagnostic accuracy of 72.5%, exceeding the performance of less experienced internists and experienced internists, though remaining below the cardiologists' benchmark of 83%. When used as a decision-support tool, GPT-4o significantly improved diagnostic accuracy in all physician groups, with the most substantial gains observed among less experienced participants. However, the findings also revealed key risks: inaccurate AI suggestions caused substantial performance deterioration in less experienced clinicians, highlighting the importance of critical evaluation and safeguards in AI integration. Overall, the study supports GPT-4o's value as an assistive tool that can elevate performance and reduce variability in clinical decision-making provided it is implemented responsibly.

6.2 FINDINGS – LIMITATIONS

The results offer a detailed view of both the opportunities and challenges of incorporating AI into diagnostic workflows. GPT-4o consistently improved diagnostic performance across all physician groups. Less experienced internists demonstrated the most substantial improvement, with a 37% increase in accuracy on difficult ECGs when supported by AI from 38% to 75%. Experienced internists also benefited from AI assistance increasing from 61% to 72%, and cardiologists, despite already high baseline performance, saw modest yet significant gains. Across all groups, physicians changed approximately one-third of their answers after seeing the AI suggestions, and in most cases these revisions led to corrected errors, indicating a net benefit from the AI input.

Nevertheless, the study also revealed a pronounced risk of overreliance. When faced with incorrect AI suggestions, less experienced internists abandoned their correct initial judgments in nearly every case, resulting in complete loss of accuracy in those instances. Experienced internists and cardiologists were less affected but still showed measurable drops in performance. This pattern reflects a clear inverse relationship between clinical experience and

susceptibility to erroneous AI input. Moreover, clinicians who reported greater familiarity with AI were less prone to being influenced by incorrect model suggestions, suggesting that trust calibration and digital literacy are essential components of safe AI use.

Several limitations must be acknowledged.

First, the controlled, questionnaire-based setting administered through a static survey platform cannot fully replicate the dynamics of real-world clinical practice, where time constraints, access to patient history, and open-ended reasoning play a significant role.

Second, the sample size of the less experienced group n = 5 was relatively small, and although the findings were striking, further validation with larger cohorts is needed.

Third, this study only tested OpenAI's models. Other AI platforms, such as Google's Gemini or Microsoft's Copilot, were not evaluated. It's possible that different models or alternative prompt designs could lead to different performance outcomes.

Fourth, the ECG cases used were adapted from curated sources, which may not fully represent the diversity and complexity encountered in daily practice. Lastly, the deliberately high proportion 20% of incorrect AI suggestions, while helpful for stress-testing the system, may not reflect typical error rates in future clinical deployment and could have biased user behavior toward either skepticism or overcorrection.

Furthermore, the study did not formally capture metrics such as decision time, confidence levels, or the rationale behind answer changes, which could provide further insights into clinician-AI interactions. Future studies incorporating such data would offer a more comprehensive understanding of user behavior and the cognitive effects of AI support.

6.3 Suggestions for Future Research

In terms of future work, several important directions emerge from this study. First, our evaluation was conducted in a controlled, questionnaire-based setting using a single-session, multiple-choice format administered through SurveyMonkey. It would be valuable to assess AI support in more realistic clinical environments, such as simulation labs, virtual consultations, or live patient care, to determine whether the effects observed in this study persist. This would also help to better understand how physicians interact with AI in dynamic, real-world contexts for example, whether they seek clarification, choose to override AI suggestions that conflict with their clinical judgment, or tend to defer to the AI when faced with uncertainty.

Second, expanding the sample size, especially among less experienced internists would improve the statistical power and generalizability of the results. While the study was adequately powered to compare cardiologists and experienced internists (with 10 participants each), the less experienced internists group was relatively small (n = 5). Although this group showed striking effects, more data is needed to confirm whether the patterns we observed, particularly around AI susceptibility and reliance, consistently hold true for early-career clinicians. A larger sample would also allow for more detailed subgroup analyses based on training background, specialty, or prior experience with AI.

Third, future work should include comparative evaluations of alternative AI models and prompt formats as the AI landscape continues to evolve rapidly. While we selected GPT-4o based on internal testing where it showed the highest accuracy on ECG questions other large multimodal models, such as updated GPT versions or Google's Gemini, are emerging and may offer different strengths. Early benchmarks suggest GPT-4o may outperform Gemini on ECG tasks, but the fast pace of AI development highlights the need for continuous comparison. In addition, exploring different ways of presenting AI output such as offering ranked differential diagnoses, probability estimates, or uncertainty flags could improve how clinicians interpret and interact with AI recommendations, potentially influencing decision-making and trust.

Finally, future work should explore ways to make AI output more transparent and trustworthy. Techniques like conformal prediction or case-based explanations could help by providing confidence levels or highlighting key features in the ECG, helping clinicians judge when to trust or question the AI. Rather than offering a single answer, AI systems could present a ranked differential or indicate uncertainty, which may help reduce overconfidence in incorrect suggestions, a key issue highlighted in this study. More broadly, our findings reinforce the need for rigorous evaluation before deploying AI in clinical workflows, ensuring that both the benefits and potential risks, such as overreliance, are fully understood and addressed.

Taken together, these directions point toward a human-centered approach for deploying diagnostic AI one that prioritizes clinical validation, user training, transparency, and thoughtful system design aligned with real-world medical practice. This study represents a step in understanding how AI models like GPT-4o can complement clinical expertise and improve diagnostic accuracy, while also highlighting the safeguards needed to make this collaboration safe, reliable, and effective.

REFERENCES

[1] J. Bajwa, U. Munir, A. Nori, and B. Williams, "Artificial intelligence in healthcare: transforming the practice of medicine," *Future Healthc J*, vol. 8, no. 2, p. e188, Jul. 2021, doi: 10.7861/FHJ.2021-0095.

[2] A. ZABOLI *et al.*, "Exploring ChatGPT's potential in ECG interpretation and outcome prediction in emergency department," *American Journal of Emergency Medicine*, vol. 88, pp. 7–11, Feb. 2025, doi: 10.1016/J.AJEM.2024.11.023.

[3] J. A. Tangsrivimol *et al.*, "Benefits, limits, and risks of ChatGPT in medicine," *Front Artif Intell*, vol. 8, p. 1518049, Jan. 2025, doi: 10.3389/FRAI.2025.1518049/BIBTEX.

[4] Y. M. Al-Worafi *et al.*, "Applications, Benefits, and Risks of ChatGPT in Medical and Health Sciences Research: An Experimental Study," *Progress In Microbes & Molecular Biology*, vol. 6, no. 1, Jun. 2023, doi: 10.36877/PMMB.A0000337.

[5] T. Dave, S. A. Athaluri, and S. Singh, "ChatGPT in medicine: an overview of its applications, advantages, limitations, future prospects, and ethical considerations," *Front Artif Intell*, vol. 6, p. 1169595, May 2023, doi: 10.3389/FRAI.2023.1169595/BIBTEX.

[6] L. Smital *et al.*, "Real-Time Quality Assessment of Long-Term ECG Signals Recorded by Wearables in Free-Living Conditions," *IEEE Trans Biomed Eng*, vol. 67, no. 10, pp. 2721–2734, Oct. 2020, doi: 10.1109/TBME.2020.2969719.

[7] A. Mariani *et al.*, "Artificial Intelligence and Its Role in the Diagnosis and Prediction of Adverse Events in Acute Coronary Syndrome: A Narrative Review of the Literature," *Life*, vol. 15, no. 4, p. 515, Apr. 2025, doi: 10.3390/LIFE15040515.

[8] B. Hunter, S. Hindocha, and R. W. Lee, "The Role of Artificial Intelligence in Early Cancer Diagnosis," *Cancers 2022, Vol. 14, Page 1524*, vol. 14, no. 6, p. 1524, Mar. 2022, doi: 10.3390/CANCERS14061524.

[9] A. Perrichot *et al.*, "Assessment of real-time electrocardiogram effects on interpretation quality by emergency physicians," *BMC Med Educ*, vol. 23, no. 1, pp. 1–7, Dec. 2023, doi: 10.1186/S12909-023-04670-X/TABLES/5.

[10] C. H. Sia *et al.*, "Fear of electrocardiogram interpretation (ECGphobia) among medical students and junior doctors," *Singapore Med J*, vol. 63, no. 12, pp. 763–768, Dec. 2022, doi: 10.11622/SMEDJ.2021078.

[11] Z. Ren, Y. Zhan, B. Yu, L. Ding, and D. Tao, "Healthcare Copilot: Eliciting the Power of General LLMs for Medical Consultation," Feb. 2024, Accessed: Jul. 01, 2025. [Online]. Available: https://arxiv.org/pdf/2402.13408

[12] G. Barabucci, V. Shia, E. Chu, B. Harack, K. Laskowski, and N. Fu, "Combining Multiple Large Language Models Improves Diagnostic Accuracy," *NEJM AI*, vol. 1, no. 11, Oct. 2024, doi: 10.1056/AIcs2400502.

[13] L. Zhu *et al.*, "Multimodal ChatGPT-4V for Electrocardiogram Interpretation: Promise and Limitations," *J Med Internet Res*, vol. 26, no. 1, 2024, doi: 10.2196/54607.

[14] G. Barabucci, V. Shia, E. Chu, B. Harack, K. Laskowski, and N. Fu, "Combining Multiple Large Language Models Improves Diagnostic Accuracy," *NEJM AI*, vol. 1, no. 11, Oct. 2024, doi: 10.1056/AIcs2400502.

[15] A. Koubaa, W. Boulila, L. Ghouti, A. Alzahem, and S. Latif, "Exploring ChatGPT Capabilities and Limitations: A Survey," *IEEE Access*, vol. 11, pp. 118698–118721, 2023, doi: 10.1109/ACCESS.2023.3326474.

[16] J. Kocoń *et al.*, "ChatGPT: Jack of all trades, master of none," *Information Fusion*, vol. 99, p. 101861, Nov. 2023, doi: 10.1016/J.INFFUS.2023.101861.

[17] S. Günay, A. Öztürk, H. Özerol, Y. Yiğit, and A. K. Erenler, "Comparison of emergency medicine specialist, cardiologist, and chat-GPT in electrocardiography assessment," *American Journal of Emergency Medicine*, vol. 80, pp. 51–60, Jun. 2024, doi: 10.1016/j.ajem.2024.03.017.

[18] T. Chanda *et al.*, "Dermatologist-like explainable AI enhances trust and confidence in diagnosing melanoma," *Nat Commun*, vol. 15, no. 1, Dec. 2024, doi: 10.1038/S41467-023-43095-4.

[19] R. Rosenbacke, Å. Melhus, and D. Stuckler, "False conflict and false confirmation errors are crucial components of AI accuracy in medical decision making," *Nat Commun*, vol. 15, no. 1, pp. 1–2, Dec. 2024, doi: 10.1038/S41467-024-50952-3;SUBJMETA=117,1634,1813,631,639,67,705;KWRD=COMPUTER+SCIENCE,MELANOMA.

[20] "150 ECG Cases - 5th Edition | Elsevier Shop." Accessed: Jul. 01, 2025. [Online]. Available: https://shop.elsevier.com/books/150-ecg-cases/hampton/978-0-7020-7458-5

[21] F. Pesapane, M. Codari, and F. Sardanelli, "Artificial intelligence in medical imaging: threat or opportunity? Radiologists again at the forefront of innovation in medicine," *Eur Radiol Exp*, vol. 2, no. 1, pp. 1–10, Dec. 2018, doi: 10.1186/S41747-018-0061-6/TABLES/2.

[22] E. Goh *et al.*, "Large Language Model Influence on Diagnostic Reasoning," *JAMA Netw Open*, vol. 7, no. 10, p. e2440969, Oct. 2024, doi: 10.1001/jamanetworkopen.2024.40969.

[23] A. J. Thirunavukarasu *et al.*, "Large language models approach expert-level clinical knowledge and reasoning in ophthalmology: A head-to-head cross-sectional study," *PLOS Digital Health*, vol. 3, no. 4, Apr. 2024, doi: 10.1371/journal.pdig.0000341.

[24] D. Prinster *et al.*, "Care to Explain? AI Explanation Types Differentially Impact Chest Radiograph Diagnostic Performance and Physician Trust in AI," *Radiology*, vol. 313, no. 2, p. e233261, Nov. 2024, doi: 10.1148/RADIOL.233261.

[25] H. H.-V. Tran *et al.*, "Electrocardiogram-Based Artificial Intelligence for Detection of Low Ejection Fraction: A Contemporary Review," *Cardiol Rev*, Jun. 2025, doi: 10.1097/CRD.0000000000000975.

[26] K. W. Johnson *et al.*, "Artificial Intelligence in Cardiology," *J Am Coll Cardiol*, vol. 71, no. 23, pp. 2668–2679, Jun. 2018, doi: 10.1016/J.JACC.2018.03.521.

[27] L. Masanneck *et al.*, "Triage Performance Across Large Language Models, ChatGPT, and Untrained Doctors in Emergency Medicine: Comparative Study," *J Med Internet Res*, vol. 26, no. 1, 2024, doi: 10.2196/53297.

[28] M. Petersen, A. Alaa, E. Kıcıman, C. Holmes, and M. van der Laan, "Artificial Intelligence–Based Copilots to Generate Causal Evidence," *NEJM AI*, vol. 1, no. 12, Nov. 2024, doi: 10.1056/AIp2400727.

[29] M. Narayan, J. Pasmore, E. Sampaio, V. Raghavan, and G. Waters, "Bias Neutralization Framework: Measuring Fairness in Large Language Models with Bias Intelligence Quotient (BiQ)," 2024.

[30] "Patient Agency and Large Language Models in."

[31] E. Goh *et al.*, "Influence of a Large Language Model on Diagnostic Reasoning: A Randomized Clinical Vignette Study.," *medRxiv*, Mar. 2024, doi: 10.1101/2024.03.12.24303785.

[32] A. Hendy *et al.*, "How Good Are GPT Models at Machine Translation? A Comprehensive Evaluation," Feb. 2023, [Online]. Available: http://arxiv.org/abs/2302.09210

[33] L. G. McCoy, A. K. Manrai, and A. Rodman, "Large Language Models and the Degradation of the Medical Record.," *N Engl J Med*, vol. 391, no. 17, pp. 1561–1564, Oct. 2024, doi: 10.1056/NEJMp2405999.

[34] M. Williams, W. Karim, J. Gelman, and M. Raza, "Ethical data acquisition for LLMs and AI algorithms in healthcare," *NPJ Digit Med*, vol. 7, no. 1, Dec. 2024, doi: 10.1038/s41746-024-01399-9.

[35] M. Tai-Seale *et al.*, "AI-Generated Draft Replies Integrated Into Health Records and Physicians' Electronic Communication," *JAMA Netw Open*, vol. 7, no. 4, pp. e246565–e246565, Apr. 2024, doi: 10.1001/JAMANETWORKOPEN.2024.6565.

[36] G. Quer and E. J. Topol, "The potential for large language models to transform cardiovascular medicine," Oct. 01, 2024, *Elsevier Ltd*. doi: 10.1016/S2589-7500(24)00151-1.

[37] E. Saveliev, J. Liů, A. Boyd, J. Liu, N. Seedat, and M. van der Schaar, "Accelerating the ML revolution in healthcare: A Human-Guided, Data-Centric Framework for LLM Co-Pilots".

[38] M. Chrab, ˛ Aszcz, K. Lorenc, and K. Seweryn, "Evaluating LLMs Robustness in Less Resourced Languages with Proxy Models," Jun. 2025, Accessed: Jul. 01, 2025. [Online]. Available: https://arxiv.org/pdf/2506.07645

[39] L. A. L. Abueg *et al.*, "The Galaxy platform for accessible, reproducible, and collaborative data analyses: 2024 update," *Nucleic Acids Res*, vol. 52, no. W1, pp. W83–W94, Jul. 2024, doi: 10.1093/NAR/GKAE410.

[40] H. Wadhwa *et al.*, "From RAGs to rich parameters: Probing how language models utilize external knowledge over parametric information for factual queries," Jun. 2024, [Online]. Available: http://arxiv.org/abs/2406.12824

[41] X. Su and Y. Gu, "Implementing Retrieval-Augmented Generation (RAG) for Large Language Models to Build Confidence in Traditional Chinese Medicine."

[42] A. Y. Hannun *et al.*, "Cardiologist-level arrhythmia detection and classification in ambulatory electrocardiograms using a deep neural network," *Nat Med*, vol. 25, no. 1, pp. 65–69, Jan. 2019, doi: 10.1038/S41591-018-0268-3;SUBJMETA=114,1305,29,631,692,699,75;KWRD=ARRHYTHMIAS,MACHINE+LEARNING.

[43] K. C. Chang *et al.*, "Usefulness of Machine Learning-Based Detection and Classification of Cardiac Arrhythmias With 12-Lead Electrocardiograms," *Canadian Journal of Cardiology*, vol. 37, no. 1, pp. 94–104, Jan. 2021, doi: 10.1016/j.cjca.2020.02.096.

[44] L. Zhu *et al.*, "Multimodal ChatGPT-4V for Electrocardiogram Interpretation: Promise and Limitations," *J Med Internet Res*, vol. 26, no. 1, 2024, doi: 10.2196/54607.

[45] F. Kücking *et al.*, "Automation Bias in AI-Decision Support: Results from an Empirical Study," *Stud Health Technol Inform*, vol. 317, pp. 298–304, Aug. 2024, doi: 10.3233/SHTI240871.

[46] H. Adam, A. Balagopalan, E. Alsentzer, F. Christia, and M. Ghassemi, "Mitigating the impact of biased artificial intelligence in emergency decision-making," *Communications Medicine*, vol. 2, no. 1, pp. 1–6, Dec. 2022, doi: 10.1038/S43856-022-00214-4;SUBJMETA=228,692,700;KWRD=HEALTH+CARE,HEALTH+SERVICES.

[47] K. Goddard, A. Roudsari, and J. C. Wyatt, "Automation bias: Empirical results assessing influencing factors," *Int J Med Inform*, vol. 83, no. 5, pp. 368–375, 2014, doi: 10.1016/j.ijmedinf.2014.01.001.

[48] S. Ackerhans, K. Wehkamp, R. Petzina, D. Dumitrescu, and C. Schultz, "Perceived Trust and Professional Identity Threat in AI-Based Clinical Decision Support Systems: Scenario-Based Experimental Study on AI Process Design Features.," *JMIR Form Res*, vol. 9, p. e64266, Mar. 2025, doi: 10.2196/64266.

[49] F. Kücking *et al.*, "Automation Bias in AI-Decision Support: Results from an Empirical Study," *Stud Health Technol Inform*, vol. 317, pp. 298–304, Aug. 2024, doi: 10.3233/SHTI240871,.

[50] "Introducing ChatGPT | OpenAI." Accessed: Jul. 01, 2025. [Online]. Available: https://openai.com/index/chatgpt/

[51] "G*Power." Accessed: Jul. 01, 2025. [Online]. Available: https://www.psychologie.hhu.de/arbeitsgruppen/allgemeine-psychologie-und-arbeitspsychologie/gpower

[52] K. C. Chang *et al.*, "Usefulness of Machine Learning-Based Detection and Classification of Cardiac Arrhythmias With 12-Lead Electrocardiograms," *Canadian Journal of Cardiology*, vol. 37, no. 1, pp. 94–104, Jan. 2021, doi: 10.1016/j.cjca.2020.02.096.

[53] S. Günay, A. Öztürk, and Y. Yiğit, "The accuracy of Gemini, GPT-4, and GPT-4o in ECG analysis: A comparison with cardiologists and emergency medicine specialists," *American Journal of Emergency Medicine*, vol. 84, pp. 68–73, Oct. 2024, doi: 10.1016/J.AJEM.2024.07.043,.

[54] H. Lee, S. Yoo, J. Kim, Y. Cho, D. Suh, and K. Lee, "A Comparative Study of Predictive model (ECG Buddy) and ChatGPT-4o for Myocardial Infarction Diagnosis via ECG image Analysis: Performance, Accuracy, and Clinical Feasibility," *medRxiv*, p. 2025.04.04.25325246, Apr. 2025, doi: 10.1101/2025.04.04.25325246.

[55] V. Pandya *et al.*, "Abstract 4142075: From GPT-4 to GPT-4o: Progress and Challenges in ECG Interpretation," *Circulation*, vol. 150, no. Suppl_1, Nov. 2024, doi: 10.1161/CIRC.150.SUPPL_1.4142075.

[56] A. M. De Roberto, F. De Marco, L. Di Biasi, D. Rossi, and G. Tortora, "Can ChatGPT-4o enhance ECG interpretation accuracy compared to cardiologists?," *Proceedings - 2024 IEEE International Conference on Bioinformatics and Biomedicine, BIBM 2024*, pp. 6852–6858, 2024, doi: 10.1109/BIBM62325.2024.10822822.

[57] Z. J. Jarou, A. Dakka, D. McGuire, and L. Bunting, "ChatGPT Versus Human Performance on Emergency Medicine Board Preparation Questions," *Ann Emerg Med*, vol. 83, no. 1, pp. 87–88, Jan. 2024, doi: 10.1016/J.ANNEMERGMED.2023.08.010.

[58] D. Prinster *et al.*, "Care to Explain? AI Explanation Types Differentially Impact Chest Radiograph Diagnostic Performance and Physician Trust in AI," *Radiology*, vol. 313, no. 2, p. e233261, Nov. 2024, doi: 10.1148/RADIOL.233261.

[59] D. Zaitchik, "When representations conflict with reality: The preschooler's problem with false beliefs and 'false' photographs," *Cognition*, vol. 35, no. 1, pp. 41–68, Apr. 1990, doi: 10.1016/0010-0277(90)90036-J.

Appendices

Appendix A. Everyday ECGs Questions Pre & Post AI suggestions

Table acronyms and their explanations:

- ca: Correct Answer Number,
- ta: Total Answer Number,
- pre: Initial physician answer without AI suggestion,
- post: Physician answer after AI suggestion,
- FALSE: The GPT model fails to predict correctly in at least 4 out of 5 attempts,
- TRUE: The GPT model succeeds in predicting correctly in at least 4 out of 5 attempts

Multiple-choice question. Bolt is the correct answer	**Less experienced Internal Medicine (ca/ta)**	**Internal Medicine (ca/ta)**	**Cardiologist (ca/ta)**	**GPT**
1. This ECG was recorded in the A&E department from a 60-year-old man who had had severe central chest pain for 1h. What is your most likely diagnosis for this patient? A. Acute lateral STEMI B. Acute inferior STEMI **C. Acute anterolateral STEMI** D. Acute anterior STEMI E. Acute posterior STEMI	4/5 pre 0/5 post	7/10 pre 4/10 post	10/10 pre 9/10 post	FALSE

Multiple-choice question. Bolt is the correct answer	**Less experienced Internal Medicine (ca/ta)**	**Internal Medicine (ca/ta)**	**Cardiologist (ca/ta)**	**GPT**
2. A 70-year-old man had had high blood pressure for many years, but it was now well controlled at 140/85. He had no symptoms, and no abnormalities were detected on physical examination. This ECG was recorded during a routine follow-up appointment. What is your most likely diagnosis for this patient? A. Dextrocardia B. RBBB and right ventricular hypertrophy C. Hypertrophic cardiomyopathy **D. Left anterior hemiblock and either left ventricular hypertrophy** E. LBBB	1/5 pre 5/5 post	8/10 pre 7/10 post	10/10 pre 10/10 post	TRUE
3. This ECG was recorded from an asymptomatic 45-year-old man at a health screening examination. What is your most likely diagnosis for this patient? **A. First-degree AV block** B. Mobitz Type I block C. Third-degree AV block D. Fourth-degree AV block (Note: There isn't a recognized fourth-degree AV block in cardiology) E. Mobitz Type II block	2/5 pre 2/5 post	08/10 pre 10/10 post	10/10 pre 10/10 post	TRUE

Multiple-choice question. Bolt is the correct answer	**Less experienced Internal Medicine (ca/ta)**	**Internal Medicine (ca/ta)**	**Cardiologist (ca/ta)**	**GPT**
4. A 60-year-old man, who 3 years earlier had had a myocardial infarction followed by mild angina, was admitted to hospital with central chest pain that had been present for 1h and had not responded to sublingual nitrates. What is your most likely diagnosis for this patient? A. Old inferior and acute inferior STEMI B. Old inferior and acute anterolateral STEMI **C. Old inferior and acute anterior STEMI** D. Non-STEMI E. Old inferior and acute posterior STEMI	3/5 pre 1/10 post	1/10 pre 1/10 post	3/10 pre 3/10 post	FALSE
5. A 60-year-old man complained of severe central chest pain, and a few minutes later became extremely breathless and collapsed. He was brought to the A&E department, where his heart rate was found to be 165 bpm, his blood pressure was unrecordable and he had signs of left ventricular failure. This is his ECG. What is your most likely diagnosis for this patient? A. Atrial fibrillation B. Ventricular fibrillation **C. Ventricular tachycardia** D. Atrial flutter E. Sinus tachycardia	2/5 pre 5/5 post	9/10 pre 10/10 post	10/10 pre 10/10 post	TRUE

Multiple-choice question. Bolt is the correct answer	**Less experienced Internal Medicine (ca/ta)**	**Internal Medicine (ca/ta)**	**Cardiologist (ca/ta)**	**GPT**
6. A 45-year-old woman had complained of occasional attacks of palpitations for 20 years, and eventually this ECG was recorded during an attack. What is your most likely diagnosis for this patient? A. Ventricular fibrillation B. Sinus tachycardia **C. AVNRT** D. Atrial fibrillation E. Normal ECG	1/5 pre 5/5 post	9/10 pre 9/10 post	10/10 pre 10/10 post	TRUE
7. A 50-year-old man is seen in the Accident and Emergency (A&E) department with severe central chest pain which has been present for 18h. What is your most likely diagnosis for this patient? **A. Acute Anterior STEMI** B. Acute Lateral STEMI C. Acute Inferior STEMI D. Acute Anterolateral STEMI E. Acute Posterior STEMI	0/5 pre 3/5 post	2/10 pre 9/10 post	8/10 pre 10/10 post	TRUE

Multiple-choice question. Bolt is the correct answer	**Less experienced Internal Medicine (ca/ta)**	**Internal Medicine (ca/ta)**	**Cardiologist (ca/ta)**	**GPT**
8. A 70-year-old man is sent to the clinic because of rather vague attacks of dizziness, which occur approximately once per week. Otherwise he is well, and there are no abnormalities on examination. What is your most likely diagnosis for this patient? A. 1st Degree AV block **B. Left anterior hemiblock and RBBB - bifascicular block** C. RBBB D. 2nd Degree AV block E. LBBB	5/5 pre 5/5 pro	2/10 pre 10/10 post	6/10 pre 9/10 post	TRUE
9. A 45-year-old man in the Coronary Care Unit with a suspected myocardial infarction suddenly becomes breathless and hypotensive, and this is his ECG. What is your most likely diagnosis for this patient? A. Ventricular tachycardia B. Ventricular fibrillation **C. Sinus tachycardia** D. LBBB E. RBBB	5/5 pre 3/5 post	7/10 pre 9/10 post	10/10 pre 9/10 post	TRUE

Multiple-choice question. Bolt is the correct answer	**Less experienced Internal Medicine (ca/ta)**	**Internal Medicine (ca/ta)**	**Cardiologist (ca/ta)**	**GPT**
10. A 20-year-old student complains of palpitations. Attacks occur about once per year. They start suddenly, his heart feels very fast and regular, and he quickly feels breathless and faint. The attacks stop suddenly after a few minutes. There are no abnormalities on examination, and this is his ECG. What is your most likely diagnosis for this patient? **A. WPW syndrome type A** B. Brugada syndrome C. Wellens syndrome D. RBBB E. LBBB	2/5 pre 5/5 post	9/10 pre 10/10 post	10/10 pre 10/10 post	TRUE
11. This ECG was recorded from a 30-year-old man who complained of chest pain: the pain did not appear to be cardiac in origin, and the physical examination was normal. What is your most likely diagnosis for this patient? A. Sinus Tachycardia **B. Normal ECG** C. Hypertrophic Cardiomyopathy D. Dextrocardia E. Dilated Cardiomyopathy	0/5 pre 0/5 post	10/10 pre 8/10 post	10/10 pre 10/10 post	FALSE

Multiple-choice question. Bolt is the correct answer	**Less experienced Internal Medicine (ca/ta)**	**Internal Medicine (ca/ta)**	**Cardiologist (ca/ta)**	**GPT**
12. A 50-year-old man is admitted to hospital as an emergency, having had chest pain characteristic of a myocardial infarction for 4h. Apart from the features associated with pain, there are no abnormal physical findings. What is your most likely diagnosis for this patient? A. Acute anterior STEMI B. Acute posterior STEMI C. Acute anterolateral STEMI **D. Acute inferior STEMI** E. Acute lateral STEMI	0/5 pre 0/5 post	10/10 pre 10/10 post	10/10 pre 10/10 post	FALSE
13. A 26-year-old woman, who has complained of palpitations in the past, is admitted to hospital via the Accident and Emergency (A&E) department with palpitations. What is your most likely diagnosis for this patient? A. Atrial fibrillation with rapid ventricular Response B. Ventricular fibrillation **C. AVNRT or AVRT** D. Atrial flutter E. Atrial fibrillation	2/5 pre 5/5 post	2/10 pre 9/10 post	9/10 pre 10/10 post	TRUE

Multiple-choice question. Bolt is the correct answer	Less experienced Internal Medicine (ca/ta)	Internal Medicine (ca/ta)	Cardiologist (ca/ta)	GPT
14. An elderly woman is admitted to hospital unconscious, evidently having had a stroke. No cardiac abnormalities are noted, but this is her ECG. What is your most likely diagnosis for this patient? A. Sinus rhythm and ventricular paced rhythm B. Ventricular fibrillation **C. Ventricular-paced rhythm and atrial fibrillation** D. Atrial flutter E. Atrial fibrillation	3/5 pre 5/5 post	9/10 pre 10/10 post	10/10 pre 10/10 post	TRUE
15. A 60-year-old woman is seen in the outpatient department, complaining of breathlessness. There are no abnormal physical findings. What is your most likely diagnosis for this patient? A. Atrial fibrillation B. Ventricular flutter C. AVNRT **D. Atrial flutter with 4:1 block** E. Ventricular fibrillation	5/5 pre 5/5 post	10/10 pre 10/10 post	9/10 pre 10/10 post	TRUE

Multiple-choice question. Bolt is the correct answer	**Less experienced Internal Medicine (ca/ta)**	**Internal Medicine (ca/ta)**	**Cardiologist (ca/ta)**	**GPT**
16. A 75-year-old woman complained of central chest discomfort on climbing hills, together with dizziness; on one occasion she had fainted while climbing stairs. What is your most likely diagnosis for this patient? **A. Sinus rhythm with LBBB** B. Sinus rhythm with RBBB C. Supraventricular tachycardia D. AVNRT E. AVRT	2/5 pre 5/5 post	9/10 pre 10/10 post	10/10 pre 10/10 post	TRUE
17. A 45-year-old man complained of palpitations, weight loss and anxiety. His blood pressure was 180/110, and his heart seemed normal. This is his ECG. His thyroid function tests, measured several times, were normal. What is your most likely diagnosis for this patient? **A. Sinus tachycardia** B. Supraventricular tachycardia C. Normal ECG D. Atrial flutter E. Atrial fibrillation	2/5 pre 0/5 post	9/10 pre 9/10 post	10/10 pre 10/10 post	FALSE

Multiple-choice question. Bolt is the correct answer	**Less experienced Internal Medicine (ca/ta)**	**Internal Medicine (ca/ta)**	**Cardiologist (ca/ta)**	**GPT**
18. This ECG was recorded from a 50-year-old man who was breathless on exertion and who had a heart murmur. What is your most likely diagnosis for this patient? **A. Left ventricular hypertrophy** B. Hypertrophic cardiomyopathy C. Dextrocardia D. Congenital heart defect E. Right ventricular hypertrophy	0/5 pre 5/5 post	1/10 pre 7/10 post	3/10 pre 9/10 post	TRUE
19. A 60-year-old man had complained of occasional episodes of palpitations for several years. Between attacks he was well, there were no physical abnormalities, and his ECG was normal. Eventually this ECG was recorded during one of his attacks. What is your most likely diagnosis for this patient? A. AVNRT B. Sinus tachycardia **C. Paroxysmal ventricular** tachycardia D. Ventricular fibrillation E. Atrial fibrillation	5/5 pre 5/5 post	2/10 pre 10/10 post	10/10 pre 10/10 post	TRUE

Multiple-choice question. Bolt is the correct answer	**Less experienced Internal Medicine (ca/ta)**	**Internal Medicine (ca/ta)**	**Cardiologist (ca/ta)**	**GPT**
20. This ECG was recorded from a 23-year-old pregnant woman who had complained of palpitations, and who had been found to have a heart murmur. What is your most likely diagnosis for this patient? A. LBBB B. Atrial fibrillation C. Ventricular fibrillation **D. RBBB and atrial extrasystoles** E. Atrial Flutter	3/5 pre 0/5 post	10/10 pre 10/10 post	10/10 pre 10/10 post	FALSE

Appendix B. More Challenging ECGs Questions Pre & Post AI suggestions

Table acronyms and their explanations:

- ca: Correct Answer Number,
- ta: Total Answer Number,
- pre: Initial physician answer without AI suggestion,
- post: Physician answer after AI suggestion,
- FALSE: The GPT model fails to predict correctly in at least 4 out of 5 attempts,
- TRUE: The GPT model succeeds in predicting correctly in at least 4 out of 5 attempts

Multiple-choice question. Bolt is the correct answer	**Less experienced Internal Medicine (ca/ta)**	**Internal Medicine (ca/ta)**	**Cardiologist (ca/ta)**	**GPT**
1. This ECG was recorded from a 30-year-old woman who complained of palpitations. What is your most likely diagnosis for this patient? A. Atrial flutter **B. Ectopic atrial beat** C. Normal EKG D. Junctional rhythm E. Atrial fibrillation	0/5 pre 0/5 post	1/10 pre 1/10 post	0/10 pre 1/10 post	FALSE

Multiple-choice question. Bolt is the correct answer	Less experienced Internal Medicine (ca/ta)	Internal Medicine (ca/ta)	Cardiologist (ca/ta)	GPT
2. A 35-year-old woman, who had had palpitations for many years without any diagnosis being made, was eventually seen in the A&E department during an attack. She looked well and was not in heart failure, and her blood pressure was 120/70. This is her ECG. What is your most likely diagnosis for this patient? A. AVNRT **B. Broad complex tachycardia with RBBB pattern, probably SVT in origin** C. Broad complex tachycardia with LBBB pattern, probably SVT in origin D. Ventricular Tachycardia E. Ventricular fibrillation	5/5 pre 5/5 post	1/10 pre 3/10 post	6/10 pre 8/10 post	TRUE
3. This ECG was recorded from a 40-year-old man who was admitted to hospital after collapsing in a supermarket. By the time he was seen he was well, and there were no abnormal physical signs. What is your most likely diagnosis for this patient? A. Wellens syndrome **B. Brugada syndrome** C. WPW syndrome D. Takatsubo syndrome E. De-Winter T waves	2/5 pre 5/5 post	8/10 pre 10/10 post	9/10 pre 10/10 post	TRUE

Multiple-choice question. Bolt is the correct answer	**Less experienced Internal Medicine (ca/ta)**	**Internal Medicine (ca/ta)**	**Cardiologist (ca/ta)**	**GPT**
4. This ECG was recorded from a 45-year-old man, who had been admitted to a coronary care unit with a myocardial infarction and who was recovering well. What is your most likely diagnosis for this patient? **A. Accelerated idioventricular rhythm** B. Junctional rhythm C. LBBB and atrial fibrillation D. RBBB and atrial fibrillation E. LBBB	0/5 pre 5/5 post	9/10 pre 10/10 post	9/10 pre 10/10 post	TRUE
5. This ECG was recorded from a healthy 25-year-old man during a routine medical examination. What is your most likely diagnosis for this patient? A. Dilated cardiomyopathy B. Hypertrophic cardiomyopathy **C. Dextrocardia** D. Normal EKG E. Brugada pattern	0/5 pre 0/5 post	1/10 pre 1/10 post	8/10 pre 8/10 post	FALSE
6. This ECG was recorded from a 15-year-old boy who collapsed while playing football but was well by the time he was seen. What is your most likely diagnosis for this patient? **A. Congenital Long QT Syndrome** B. Hypertrophic Cardiomyopathy C. Normal EKG D. Structural heart disease E. Congenital Short QT Syndrome	1/5 pre 5/5 post	1/10 pre 3/10 post	3/10 pre 5/10 post	TRUE

Multiple-choice question. Bolt is the correct answer	**Less experienced Internal Medicine (ca/ta)**	**Internal Medicine (ca/ta)**	**Cardiologist (ca/ta)**	**GPT**
7. A 35-year-old woman, who had had attacks of what sounded like a paroxysmal tachycardia for many years, was seen in the A&E department, and this ECG was recorded. What is your most likely diagnosis for this patient? A. AVNRT **B. SVT and Type B WPW syndrome** C. Type A WPW syndrome D. Aberrantly conducted SVT E. Atrial fibrillation with rapid ventricular Response	3/5 pre 0/5 post	2/10 pre 0/10 post	5/10 pre 2/10 post	FALSE
8. This ECG was recorded from a 50-year-old man admitted to hospital following 2h of central chest pain that was characteristic of a myocardial infarction. His ECG had been normal 6 months ago. What is your most likely diagnosis for this patient? **A. LBBB and ventricular extrasystoles** B. RBBB and ventricular extrasystoles C. LBBB and AVNRT D. WPW syndrome E. Brugada syndrome	0/5 pre 5/5 post	7/10 pre 10/10 post	9/10 pre 9/10 post	TRUE

Multiple-choice question. Bolt is the correct answer	**Less experienced Internal Medicine (ca/ta)**	**Internal Medicine (ca/ta)**	**Cardiologist (ca/ta)**	**GPT**
9. A 30-year-old man, who had complained of palpitations for many years without anything abnormal being found, came to the A&E department during an attack, and this ECG was recorded. Apart from signs of marked anxiety, there were no unusual findings except a heart rate of 140 bpm. What is your most likely diagnosis for this patient? A. Atrial flutter B. AVRT C. Atrial fibrillation **D. Atrial tachycardia** E. Sinus tachycardia	1/5 pre 0/5 post	2/10 pre 1/10 post	0/10 pre 1/10 post	FALSE
10. A 70-year-old woman, admitted to hospital because of increasing heart failure of uncertain cause, collapsed and was found to have a very rapid pulse and a low blood pressure. This is her ECG. She recovered spontaneously. What is your most likely diagnosis for this patient? A. Ventricular fibrillation B. LBBB C. Atrial fibrillation **D. Ventricular Tachycardia** E. Atrial fibrillation and RBBB	0/5 pre 5/5 post	10/10 pre 10/10 post	10/10 pre 10/10 post	TRUE

Multiple-choice question. Bolt is the correct answer	**Less experienced Internal Medicine (ca/ta)**	**Internal Medicine (ca/ta)**	**Cardiologist (ca/ta)**	**GPT**
11. A 40-year-old man is seen in the outpatient department with a history that suggests a myocardial infarction 3 weeks previously. There are no abnormalities on examination, and this is his ECG. There are two possible explanations for the abnormality it shows, though only one of these would explain his history. What is your most likely diagnosis for this patient? A. Lateral STEMI B. RBBB **C. Probable posterior myocardial infarction** D. Right ventricular hypertrophy E. Anterior myocardial infarction	4/5 pre 5/5 post	6/10 pre 9/10 post	9/10 pre 10/10 post	TRUE
12. This ECG was recorded as part of a routine examination of a healthy 25-year-old professional athlete. There were no abnormal physical findings. What is your most likely diagnosis for this patient? A. Previous lateral MI **B. Probable hypertrophic cardiomyopathy** C. Non-STEMI D. Incomplete LBBB E. RBBB	4/5 pre 0/5 post	10/10 pre 10/10 post	10/10 pre 10/10 post	FALSE

Multiple-choice question. Bolt is the correct answer	**Less experienced Internal Medicine (ca/ta)**	**Internal Medicine (ca/ta)**	**Cardiologist (ca/ta)**	**GPT**
13. This ECG was recorded from a 30-year-old woman admitted to hospital with diabetic ketoacidosis. What is your most likely diagnosis for this patient? A. Hypokalemia **B. Hyperkalemia** C. Hyperphosphatemia D. Hypercalcemia E. Hypocalcemia	2/5 pre 5/5 post	10/10 pre 10/10 post	10/10 pre 10/10 post	TRUE
14. This ECG was recorded from a 37-year-old man admitted to hospital for a routine orthopaedic operation. What is your most likely diagnosis for this patient? A. **Normal ECG** B. Atrial flutter C. Atrial fibrillation D. Ventricular fibrill E. Nodal rhythm	0/5 pre 5/5 post	9/10 pre 10/10 post	10/10 pre 10/10 post	TRUE
15. This ECG was recorded from a 30-year-old woman with severe rheumatoid arthritis, who was admitted to hospital with central chest pain. She was a non-smoker and had no risk factors for coronary artery disease. What is your most likely diagnosis for this patient? A. STEMI **B. Pericarditis** C. Hypertrophic cardiomyopathy D. Dilated cardiomyopathy E. LBBB	2/5 pre 5/5 post	10/10 pre 10/10 post	10/10 pre 10/10 post	TRUE

Multiple-choice question. Bolt is the correct answer	Less experienced Internal Medicine (ca/ta)	Internal Medicine (ca/ta)	Cardiologist (ca/ta)	GPT
16. This ECG was recorded from a 15-year-old boy who collapsed while playing football. His brother had died suddenly. What is your most likely diagnosis for this patient? A. Congenital heart disease B. Sinus rhythm **C. Long QT syndrome** D. Short QT syndrome E. WPW syndrome	2/5 pre 5/5 post	5/10 pre 10/10 post	10/10 pre 10/10 post	TRUE
17. An elderly man was admitted unconscious after a stroke, and this was his ECG. What is your most likely diagnosis for this patient? **A. Dual chamber pacemaker** B. 2nd degree AV block C. Atrial paced rhythm D. 3rd degree complete AV block E. 1st degree AV block	0/5 pre 5/5 post	10/10 pre 10/10 post	10/10 pre 10/10 post	TRUE
18. A 70-year-old man with long-standing high blood pressure has had attacks of dizziness over several weeks. His pulse feels irregular, but there are no other abnormal signs. This was his ECG. What is your most likely diagnosis for this patient? A. WPW syndrome B. Wenckebach type 2 block **C. Second degree block of both the Wenckebach type and Mobitz type 2, and also first degree block** D. Mobitz type 2 block E. Wellens Syndrome	5/5 pre 5/5 post	1/10 pre 6/10 post	2/10 pre 7/10 post	TRUE

Multiple-choice question. Bolt is the correct answer	Less experienced Internal Medicine (ca/ta)	Internal Medicine (ca/ta)	Cardiologist (ca/ta)	GPT
19. A 35-year-old white man is seen in the outpatient department complaining of chest pain on exertion, sometimes with exertion-induced dizziness, and this is his ECG. What is your most likely diagnosis for this patient? A. Left anterior hemiblock B. RBBB **C. Hypertrophic cardiomyopathy** D. LBBB E. RBBB and left anterior hemiblock	3/5 pre 5/5 post	9/10 pre 10/10 post	10/10 pre 10/10 post	TRUE
20. This ECG was recorded from a 75-year-old woman who complained of attacks of dizziness. What is your most likely diagnosis for this patient? A. LBBB B. Normal ECG **C. Sinus rhythm and first-degree block** D. RBBB E. Second-degree AV block	4/5 pre 5/5 post	10/10 pre 10/10 post	10/10 pre 10/10 post	TRUE

Appendix C. Fake ECGs Questions Pre & Post AI suggestions

Table acronyms and their explanations:

- ca: Correct Answer Number,
- ta: Total Answer Number,
- pre: Initial physician answer without AI suggestion,
- post: Physician answer after AI suggestion,

Multiple-choice question. Bolt is the correct answer	**Less experienced Internal Medicine (ca/ta)**	**Internal Medicine (ca/ta)**	**Cardiologist (ca/ta)**
1. This ECG was recorded as part of the routine investigation of a 40-year-old man who was admitted to hospital following a first seizure. He was unconscious and had a stiff neck and bilateral extensor plantar responses. His heart was clinically normal. What is your most likely diagnosis for this patient? A. Anterolateral ischemia B. Normal ECG C. Pericarditis **D. Anterolateral T wave inversion due to subarachnoid haemorrhage** E. Left ventricular hypertrophy	2/5 pre 0/5 post	9/10 pre 8/10 post	10/10 pre 10/10 post

Multiple-choice question. Bolt is the correct answer	**Less experienced Internal Medicine (ca/ta)**	**Internal Medicine (ca/ta)**	**Cardiologist (ca/ta)**
2. An 18-year-old student complains of occasional attacks of palpitations. These start suddenly without provocation; the heartbeat seems regular and is 'too fast to count'. During attacks she does not feel dizzy or breathless, and the palpitations stop suddenly after a few seconds. Physical examination is normal, and this is her ECG. What is your most likely diagnosis for this patient? **A. Possible pre-excitation** B. Sinus tachycardia C. AVNRT D. Sporadic ventricular extrasystoles E. Atrial fibrillation	2/5 pre 0/5 post	9/10 pre 5/10 post	10/10 pre 10/10 post
3. A 30-year-old man, who had had attacks of palpitations for several years, was seen during an attack, and this ECG was recorded. He was breathless and his blood pressure was unrecordable. What is your most likely diagnosis for this patient? A. Atrial fibrillation with rapid ventricular response B. Ventricular fibrillation **C. Ventricular tachycardia** D. Supraventricular tachycardia E. Sinus tachycardia with aberration	0/5 pre 0/5 post	10/10 pre 9/10 post	10/10 pre 10/10 post

Multiple-choice question. Bolt is the correct answer	**Less experienced Internal Medicine (ca/ta)**	**Internal Medicine (ca/ta)**	**Cardiologist (ca/ta)**
4. This ECG was recorded from an 80-year-old woman who had been found unconscious with physical signs suggesting a stroke. What is your most likely diagnosis for this patient? A. Atrial flutter and hypotension B. Sinus bradycardia and hypothermia **C. Atrial fibrillation and hypothermia** D. Ventricular fibrillation and hyperkalemia E. Second-degree AV block	0/5 pre 0/5 post	0/10 pre 0/10 post	0/10 pre 0/10 post
5. A 25-year-old woman, who had had episodes of what sound like a paroxysmal tachycardia for 10 years, produced this ECG when seen during an attack. What is your most likely diagnosis for this patient? **A. Atrial fibrillation and the WPW syndrome type A** B. Atrial flutter with RBBB C. Ventricular tachycardia D. Sinus tachycardia with PVCs E. AVNRT with LBBB	3/5 pre 0/5 post	8/10 pre 8/10 post	10/10 pre 10/10 post

Multiple-choice question. Bolt is the correct answer	**Less experienced Internal Medicine (ca/ta)**	**Internal Medicine (ca/ta)**	**Cardiologist (ca/ta)**
6. A 75-year-old woman was admitted with heart failure. She had been treated with digoxin, ramipril, Frumil and Spironolactone. This is her ECG. What is your most likely diagnosis for this patient? A. Hypokalemia with LBBB B. Digoxin toxicity C. Atrial fibrillation with LVH D. True posterior myocardial infarction **E. Hyperkalaemia and RBBB**	0/5 pre 0/5 post	2/10 pre 0/10 post	9/10 pre 2/10 post
7. This ECG was recorded as part of the health screening of an asymptomatic 40-year-old man. What is your most likely diagnosis for this patient? A. Sinus rhythm and intermittent pacemaker rhythm B. Sinus rhythm with frequent PVCs **C. Sinus rhythm and accelerated idioventricular rhythm** D. Sinus rhythm with RBBB E. Sinus rhythm and atrial flutter	3/5 pre 0/5 post	4/10 pre 4/10 post	8/10 pre 8/10 post

Multiple-choice question. Bolt is the correct answer	**Less experienced Internal Medicine (ca/ta)**	**Internal Medicine (ca/ta)**	**Cardiologist (ca/ta)**
8. A 45-year-old man was admitted to hospital with a history of 2h of ischaemic chest pain. His blood pressure was 150/80, and there were no signs of heart failure. This was his ECG. What is your most likely diagnosis for this patient? A. Atrial flutter B. Supraventricular tachycardia C. Wolff-Parkinson-White syndrome **D. Broad complex tachycardia – probably ventricular tachycardia** E. Sinus tachycardia	5/5 pre 0/5 post	2/10 pre 2/10 post	4/10 pre 4/10 post
9. A 30-year-old woman, who had been treated for depression for several years, was admitted to hospital as an emergency following deliberate self-harm involving a small number of aspirin tablets. There were no abnormalities on examination, but this was her ECG. What is your most likely diagnosis for this patient? A. Anterolateral ischemia **B. Anterolateral T wave inversion due to lithium therapy** C. Normal ECG D. Ventricular hypertrophy E. Pericarditis	3/5 pre 0/5 post	9/10 pre 9/10 post	10/10 pre 10/10 post

Multiple-choice question. Bolt is the correct answer	**Less experienced Internal Medicine (ca/ta)**	**Internal Medicine (ca/ta)**	**Cardiologist (ca/ta)**
10. This ECG was recorded in the A&E department from a 25-year-old man with severe chest pain. No physical abnormalities had been detected. What is your most likely diagnosis for this patient? A. Acute anterior STEMI **B. Widespread ST segment elevation, suggesting pericarditis** C. Early repolarization D. Left ventricular hypertrophy E. RBBB	5/5 pre 0/5 post	8/10 pre 6/10 post	9/10 pre 10/10 post

Appendix D. Ethical Approval of the Research Protocol (Greek Version)

ΑΠΟΦΑΣΗ

ΕΠΙΤΡΟΠΗΣ ΗΘΙΚΗΣ ΚΑΙ ΔΕΟΝΤΟΛΟΓΙΑΣ ΤΗΣ ΕΡΕΥΝΑΣ (Ε.Η.Δ.Ε.) ΤΟΥ ΕΘΝΙΚΟΥ ΜΕΤΣΟΒΙΟΥ ΠΟΛΥΤΕΧΝΕΙΟΥ

ΓΙΑ

ΕΓΚΡΙΣΗ ΕΡΕΥΝΗΤΙΚΟΥ ΠΡΩΤΟΚΟΛΛΟΥ

ΕΜΠΙΣΤΕΥΤΙΚΟ ΕΓΓΡΑΦΟ

1/3

Τίτλος μελέτης για την οποία ζητήθηκε έγκριση
«Synergistic AI-Human Integration for Cardiology»
Επιστημονικός Υπεύθυνος της μελέτης
Κωνσταντίνα Νικήτα (Καθηγήτρια ΕΜΠ)
Είδος προτεινομένης μελέτης
Η παρούσα έρευνα διερευνά τη δυνατότητα ενσωμάτωσης της Τεχνητής Νοημοσύνης (ΤΝ) στη διαδικασία διάγνωσης καρδιολογικών περιστατικών. Στόχος είναι να αξιολογηθεί αν η χρήση ενός συστήματος ΤΝ (Μεγάλα Γλωσσικά Μοντέλα) μπορεί να ενισχύσει την ικανότητα των γιατρών στη λήψη αποφάσεων. Οι συμμετέχοντες καρδιολόγοι και παθολόγοι καλούνται να αναλύσουν περιπτώσεις ηλεκτροκαρδιογραφημάτων και να προτείνουν διάγνωση. Στη συνέχεια, λαμβάνουν υποδείξεις από την ΤΝ και επανεξετάζουν τις απαντήσεις τους. Μέσω αυτής της διαδικασίας καταγράφονται οι αλλαγές στις διαγνωστικές εκτιμήσεις, ενώ αξιολογείται και η κριτική στάση των γιατρών απέναντι σε σωστές και ελεγχόμενα λανθασμένες προτάσεις της ΤΝ. Τα δεδομένα θα αναλυθούν για την κατανόηση της συμβολής της ΤΝ στη βελτίωση της διαγνωστικής ικανότητας. Η έρευνα στοχεύει στη διαμόρφωση ενός συνεργατικού μοντέλου ανθρώπου-μηχανής που θα ενισχύσει την κλινική πρακτική και θα προάγει τη χρήση καινοτόμων τεχνολογιών στη διάγνωση καρδιαγγειακών παθήσεων.
Αριθμός Πρωτοκόλλου Ε.Η.Δ.Ε./ Αριθμός Πρωτοκόλλου Ε.Α.Κ.Ε.
2008/08.01.2025
Αριθμός & Ημερομηνία Απόφασης Επιτροπής Ηθικής και Δεοντολογίας της Έρευνας (Ε.Η.Δ.Ε.)
Συνεδρίαση 14.01.2025, Θέμα 1.3
Απόφαση Επιτροπής Ηθικής και Δεοντολογίας της Έρευνας (Ε.Η.Δ.Ε.)
Εγκρίνεται
Μέλη της Επιτροπής
Α. Ανδρικόπουλος (Πρόεδρος), Δ. Μαμάης, Ε. Κορωνάκη, Μ. Δαμανάκη, Κ. Κορδάτος, Χ. Κούβαρης, Ε. Στάη, Β. Κορμπάκη, Ι. Λαδάς
Σχόλια από την Επιτροπή Ηθικής και Δεοντολογίας της Έρευνας (Ε.Η.Δ.Ε.) με βάση τα οποία λήφθηκε η απόφαση για την αίτηση που υποβλήθηκε
Μελετώντας το ερευνητικό πρωτόκολλο και όλα τα σχετικά δικαιολογητικά/πρόσθετες εγκρίσεις, όπως κατατέθηκαν στην Επιτροπή Ηθικής και Δεοντολογίας της Έρευνας (Ε.Η.Δ.Ε.), και λαμβάνοντας υπόψη τους σκοπούς και τα αναμενόμενα οφέλη, τη μεθοδολογία της

2/3

έρευνας, την έλλειψη σύγκρουσης συμφερόντων από τους ερευνητές και την έλλειψη πιθανών κινδύνων για τα υποκείμενα της έρευνας, σύμφωνα με τα διαλαμβανόμενα στη σχετική εισηγητική έκθεση,

η ΕΗΔΕ

διαπιστώνει και ομόφωνα εγκρίνει την υποβληθείσα αίτηση (άρθρο 279 παρ. 1 ν. 4957)

Η παρούσα απόφαση της ΕΠΙΔΕ σε καμία περίπτωση ΔΕΝ υποκαθιστά την απαιτούμενη από άλλη αρμόδια δημόσια υπηρεσία, διοικητικό όργανο ή ανεξάρτητη διοικητική Αρχή, έγκριση ή αδειοδότηση του παρόντος ερευνητικού έργου/μελέτης που δύναται επιπλέον να απαιτείται εκ του νόμου.

Ημερομηνία έκδοσης απόφασης		
Έτος: **2025**	Μήνας: **Ιανουάριος**	Ημέρα: **14, Τρίτη**

Υπογράφει ο Πρόεδρος της Επιτροπής			
Θέση	Όνομα	Επώνυμο	Υπογραφή
Πρόεδρος	Ανδρέας	Ανδρικόπουλος	

3/3

Appendix E. Ethical Approval of the Research Protocol (English Version)

DECISION

OF THE RESEARCH ETHICS AND DEONTOLOGY COMMITTEE (R.E.D.C.) OF THE NATIONAL TECHNICAL UNIVERSITY OF ATHENS

FOR

THE APPROVAL OF A RESEARCH PROTOCOL

CONFIDENTIAL DOCUMENT

1/3

Title of the study for which approval was requested
«Synergistic AI-Human Integration for Cardiology»
Scientific Supervisor of the study
Konstantina Nikita (Professor NTUA)
Type of proposed study
This study investigates the potential integration of Artificial Intelligence (AI) into the diagnostic process of cardiological cases. The aim is to assess whether the use of an AI system (Large Language Models) can enhance physicians' decision-making capabilities. Participating cardiologists and internists are asked to analyze electrocardiogram cases and provide a diagnosis. They then receive suggestions from the AI and are prompted to re-evaluate their initial responses. This process records any changes in diagnostic assessments, while also evaluating the physicians' critical stance toward both correct and deliberately incorrect AI-generated suggestions. The data will be analyzed to understand the contribution of AI to improving diagnostic accuracy. The study aims to develop a collaborative human–machine model that will support clinical practice and promote the use of innovative technologies in the diagnosis of cardiovascular diseases.
Protocol Number assigned by the Research Ethics and Deontology Committee (R.E.D.C.)/ Special Research Fund (S.R.F.)
2008/08.01.2025
Number and date of the decision by the Research Ethics and Deontology Committee (R.E.D.C.)
Meeting held on 14 January 2025, Agenda Item 1.3
Decision by the Research Ethics and Deontology Committee (R.E.D.C.)
Approved
Committee members
A. Andreopoulos (Director), D. Mamais, E.Koronaki, M. Damanaki, K. Kordatos, C. Kouvaris, E. Stai, V. Kormpakis, I. Ladas
Comments by the Research Ethics and Deontology Committee (R.E.D.C.) that formed the basis for the decision on the submitted application
Having reviewed the research protocol and all relevant supporting documents/additional approvals, as submitted to the Research Ethics and Deontology Committee (R.E.D.C.), and taking into consideration the objectives and the anticipated benefits, as well as the methodology of

2/3

the study, the absence of conflicts of interest by the researchers, and the absence of potential risks to the research subjects, in accordance with the statements contained in the relevant explanatory report,

The R.E.D.C.

confirms and unanimously approves the submitted application (Article 279, par. 1 Law 4957)

This decision of the R.E.D.C. in no case substitutes for any approval or license that may additionally be required by law from another competent public authority, administrative body, or independent administrative authority, for the present research project/study.

Date of decision issuance			
Year: 2025	**Month: January**		**Day: 14, Tuesday**
Signed by the Committee Chair			
Position	Name	Surname	Signature
Director	Andreas	Andreopoulos	

3/3

Appendix F. Study Protocol (Greek Version)

Σύντομη Περιγραφή Έρευνας

Θεωρητικό Υπόβαθρο

Η Τεχνητή Νοημοσύνη (ΤΝ) αναδιαμορφώνει τον τομέα των διαγνωστικών πρακτικών, προσφέροντας νέα εργαλεία που μπορούν να ενισχύσουν την ποιότητα και την ταχύτητα των ιατρικών αποφάσεων. Ειδικότερα στην καρδιολογία, όπου η ακριβής και άμεση διάγνωση των καρδιαγγειακών παθήσεων είναι κρίσιμη για την έκβαση της θεραπείας, η ενσωμάτωση προηγμένων συστημάτων ΤΝ ανοίγει νέους ορίζοντες. Τα σύγχρονα συστήματα που αξιοποιούν Μεγάλα Γλωσσικά Μοντέλα (LLMs) με τεχνολογία Αναγνώρισης Ανάκτησης (RAG) επιτρέπουν στους καρδιολόγους να αποκτούν πρόσβαση σε τεράστιο όγκο εξειδικευμένης ιατρικής γνώσης σε πραγματικό χρόνο, προσφέροντας διαγνωστικές προτάσεις με βάση τα μοναδικά δεδομένα κάθε ασθενούς.

Τα LLMs, με την τεχνολογία RAG, λειτουργούν ως «ψηφιακοί συνεργάτες» ή AI co-pilots, οι οποίοι παρέχουν προτάσεις βασισμένες σε ατομικά και συγκριτικά δεδομένα ασθενών. Αυτή η υποστήριξη διευκολύνει τους γιατρούς στη διαδικασία λήψης αποφάσεων, επιτρέποντάς τους να εντοπίζουν πιθανές διαγνώσεις και να επανεκτιμούν τις αρχικές τους εκτιμήσεις. Χάρη στη δυνατότητα ενσωμάτωσης πληροφοριών από προηγούμενα ιατρικά δεδομένα, τα συστήματα αυτά βοηθούν στην παροχή αναλυτικών και ακριβών προτάσεων, ενισχύοντας την κλινική λογική και την αυτοπεποίθηση των ιατρών.

Παρά τις δυνατότητές τους, τα συστήματα αυτά δεν αντικαθιστούν τη γνώση και την εμπειρία του γιατρού, αλλά αντίθετα λειτουργούν συνεργατικά, επιτρέποντας στους γιατρούς να διατηρούν τον έλεγχο της διαδικασίας διάγνωσης. Αυτή η διαδραστική συνεργασία ανθρώπου και ΤΝ δημιουργεί ένα νέο πρότυπο στην κλινική πράξη, όπου η υποστήριξη της ΤΝ αξιοποιείται σε συνδυασμό με την κρίση και την εμπειρία του γιατρού, διαμορφώνοντας ένα περιβάλλον αλληλεπίδρασης ανθρώπου-μηχανής που προάγει την εμπιστοσύνη.

Ωστόσο, η εφαρμογή της ΤΝ στην ιατρική παραμένει ένα αναδυόμενο πεδίο με πολλές ανοιχτές προκλήσεις. Θέματα όπως η αντικειμενικότητα των διαγνωστικών προτάσεων, η αποδοχή των συστημάτων αυτών από τους γιατρούς, η κριτική τους διάθεση απέναντί τους, καθώς και η βιωσιμότητα της συνεργασίας τους σε πραγματικά κλινικά περιβάλλοντα, χρειάζονται περαιτέρω διερεύνηση. Η ανάγκη για εμπεριστατωμένη αξιολόγηση αυτών των συστημάτων είναι καθοριστική, καθώς θα βοηθήσει να κατανοήσουμε πώς η αλληλεπίδραση ανθρώπου-ΤΝ μπορεί να ενισχύσει την ακρίβεια και την αποτελεσματικότητα στη διάγνωση, συμβάλλοντας στην ομαλή ενσωμάτωσή τους στην καθημερινή ιατρική πρακτική και εξασφαλίζοντας την ποιοτική φροντίδα των ασθενών.

Ερευνητική Διαδικασία

Η έρευνα επικεντρώνεται στην αλληλεπίδραση καρδιολόγων και παθολόγων με ένα σύστημα Τεχνητής Νοημοσύνης (ΤΝ) που υποστηρίζει τη διαγνωστική διαδικασία. Στόχος είναι η αξιολόγηση του βαθμού στον οποίο η ενσωμάτωση της ΤΝ μπορεί να βελτιώσει την ακρίβεια και την αυτοπεποίθηση των γιατρών στη διάγνωση καρδιολογικών περιστατικών, καθώς και στην αξιολόγηση της εμπιστοσύνης και της κριτικής στάσης των ιατρών απέναντι στην ΤΝ. Οι

συμμετέχοντες θα αξιολογήσουν κλινικές περιπτώσεις με και χωρίς την υποστήριξη της ΤΝ, μέσω συμπλήρωσης ενός ερωτηματολογίου. Η διαδικασία θα έχει διάρκεια περίπου 50-60 λεπτά συνολικά. Η παρούσα έρευνα πραγματοποιείται στο πλαίσιο δύο μεταπτυχιακών διπλωματικών εργασιών για το ΔΠΜΣ Μεταφραστική Βιοϊατρική Μηχανική και Επιστήμη.

Η παρούσα έρευνα περιλαμβάνει τα εξής στάδια:

1. **Επιλογή και Περιγραφή Συμμετεχόντων**:
    - Η έρευνα απευθύνεται σε καρδιολόγους και παθολόγους. Η προσέγγιση των συμμετεχόντων γιατρών πραγματοποιείται μέσω αναζήτησης στο ακαδημαϊκό, επαγγελματικό και προσωπικό περιβάλλον. Οι γιατροί που πληρούν τα κριτήρια συμμετοχής ενημερώνονται προσωπικά για τον σκοπό και τη διαδικασία της έρευνας. Η ενημέρωση περιλαμβάνει αναλυτική επεξήγηση του στόχου της μελέτης, της φύσης της αλληλεπίδρασής τους με το σύστημα Τεχνητής Νοημοσύνης, καθώς και των πιθανών ωφελειών και κινδύνων από τη συμμετοχή τους. Κατά την έναρξη κάθε συνεδρίας, ο συμμετέχον δηλώνει αν σε ποια ειδικότητα ανήκει, την εργασιακή του εμπειρία και κατά πόσο ολοκλήρωσε την ειδικότητά του στην Ελλάδα. Επιπλέον, δηλώνει το επίπεδο εξοικείωσής τους με εφαρμογές της ΤΝ.
2. **Παρουσίαση Κλινικών Περιστατικών**:
    - Το ερωτηματολόγιο που καλούνται να συμπληρώσουν οι συμμετέχοντες αποτελείται από κλινικά περιστατικά που περιλαμβάνουν δεδομένα ηλεκτροκαρδιογραφημάτων (ECG) μαζί με βασικά δημογραφικά στοιχεία του ασθενούς (ηλικία, φύλο, συμπτώματα). Τα περιστατικά αυτά έχουν αντληθεί από εγχειρίδιο εκπαίδευσης καρδιολόγων [1]. Τα περιστατικά έχουν μετατραπεί σε ερωτήσεις πολλαπλής επιλογής σε προηγούμενη έρευνα [2]. Για τις ανάγκες τις παρούσας έρευνας έχουν χρησιμοποιηθεί επιπλέον 10 ερωτήσεις του εγχειριδίου, οποίες μετατράπηκαν σε πολλαπλής επιλογής με τη βοήθεια καρδιολόγου. Σκοπός της προσθήκης των ερωτήσεων αυτών όπως περιγράφεται αναλυτικότερα και παρακάτω, είναι η παρατήρηση και η ανάλυση της στάσης και της κρίσης των ιατρών απέναντι σε ερωτήσεις όπου οι πρόταση του AI δεν είναι έγκυρη.
3. **Αρχική Διάγνωση**:
    - Οι συμμετέχοντες εξετάζουν κάθε περίπτωση ξεχωριστά και καταγράφουν την αρχική τους διάγνωση με βάση τα προσφερόμενα δεδομένα ECG και τη δική τους κλινική εμπειρία.
4. **Παρέμβαση και Υπόδειξη από την ΤΝ**:
    - Μετά την καταγραφή της αρχικής διάγνωσης, το ίδιο περιστατικό παρουσιάζεται εκ νέου με μια πρόταση διάγνωσης από την ΤΝ. Η πρόταση αυτή έχει δημιουργηθεί από ένα LLM σύστημα. Οι γιατροί μπορούν να επανεξετάσουν την αρχική τους εκτίμηση, χρησιμοποιώντας την πρόταση της ΤΝ ως οδηγό, ή να επιβεβαιώσουν την αρχική τους απόφαση.

- Ανάμεσα στις προτάσεις του μοντέλου ΤΝ, θα υπάρχουν και κάποιες (~20% του συνόλου) οι οποίες ελεγχόμενα και στοχευμένα θα είναι λανθασμένες (χωρίς την πρότερη ενημέρωση του ιατρού). Με τη βοήθεια των προτάσεων αυτών σκοπός είναι να καταγραφεί και να αναλυθεί η επιδραστικότητα της ΤΝ σε ιατρούς διαφόρων επιπέδων εξειδίκευσης αλλά και εξοικείωσης με εφαρμογές ΤΝ, καθώς και η συμπεριφορά αυτών μπροστά σε περιστατικά όπου το γνωστικό επίπεδο ενός μοντέλου είναι αναντίστοιχο των γνώσεων του ιατρού. Η παραπάνω διαδικασία σχεδιάζεται και εφαρμόζεται με τέτοιο τρόπο ώστε να μην επηρεάσει τις περιπτώσεις όπου η αλληλεπίδραση του ιατρού γίνεται με τις πραγματικές προτάσεις του μοντέλου ΤΝ.

5. **Ολοκλήρωση συνεδρίας:**

   - Με το πέρας της συμπλήρωσης του ερωτηματολογίου για τον κάθε ιατρό, πραγματοποιείται ενημέρωση του συμμετέχοντα ως προς την ύπαρξη των λανθασμένων προτάσεων με σκοπό να κατανοήσουμε την εμπειρία του ιατρού κατά την απάντηση των ερωτήσεων και κατά την αλληλεπίδραση του με τις (σωστές και λανθασμένες) προτάσεις της ΤΝ.

6. **Συλλογή και Καταγραφή Δεδομένων**:

   - Το ονοματεπώνυμο κάθε συμμετέχοντα θα εμφανίζεται μόνο στην υπογραφή του στο έντυπο συγκατάθεσης, το οποίο δεν θα ψηφιοποιηθεί και δεν θα συνδέεται με τα υπόλοιπα δεδομένα που συλλέγονται. Παράλληλα, για να καταστεί το δικαίωμα των συμμετεχόντων να αποσύρουν τα δεδομένα τους από την έρευνα εφικτό, θα τους δίνεται κατά τη στιγμή της έναρξης της διαδικασίας ένας τυχαίος κωδικός, ο οποίος θα καταγράφεται από τους ίδιους στο ερωτηματολόγιο τους. Η σύνδεση ονόματος και κωδικού θα είναι δυνατή μόνο από το συμμετέχοντα, ο οποίος στην περίπτωση που επιθυμεί να αποσύρει τις απαντήσεις του με τον κωδικό του επιτρέπει στους ερευνητές/φοιτητές να ταυτοποιήσουν τις απαντήσεις του.

   - Πέρα από το ονοματεπώνυμο δεν θα καταγράφονται άλλα προσωπικά στοιχεία. Όλα τα δεδομένα που θα συλλεχθούν στο πλαίσιο της έρευνας θα παραμείνουν πλήρως ανώνυμα. Για τις ανάγκες της έρευνας θα καταγράφεται η ειδικότητα των συμμετεχόντων (καρδιολόγος ή παθολόγος), η επαγγελματική τους εμπειρία (έτη από την απόκτηση ειδικότητας), κατά πόσο ολοκλήρωσαν την ειδικότητά τους στην Ελλάδα και το επίπεδο εξοικείωσής τους με εφαρμογές τεχνητής νοημοσύνης.

   - Όλα τα ανώνυμα δεδομένα της έρευνας συλλέγονται μέσω του ερωτηματολογίου το οποίο έχει υλοποιηθεί στην υπηρεσία SurveyMonkey [3]. Με την ολοκλήρωση της διαδικασίας συλλογής των δεδομένων της έρευνας οι ανώνυμες απαντήσεις θα διαγραφούν από την υπηρεσία SurveyMonkey και θα αποθηκευτούν για την ανάλυση τους . Σύμφωνα με τους όρους χρήσης της υπηρεσίας SurveyMonkey, η διαγραφή των απαντήσεων είναι οριστική, η πρόσβαση σε αυτές χάνεται άμεσα και η

οριστική τους διαγραφή από τα συστήματα της υπηρεσίας γίνεται σε 60 μέρες.

- Η πρόσβαση στα δεδομένα θα περιοριστεί αυστηρά στους υπεύθυνους της έρευνας, ενώ δεν θα χρησιμοποιηθούν για άλλους σκοπούς πέραν της παρούσας μελέτης. Όσον αφορά πιθανά πνευματικά δικαιώματα των συμμετεχόντων, οι πληροφορίες που θα συλλεχθούν δεν περιλαμβάνουν δεδομένα που θα μπορούσαν να ταυτοποιήσουν κάποιο άτομο, διασφαλίζοντας έτσι ότι δεν τίθεται ζήτημα παραβίασης των δικαιωμάτων τους. Επιπλέον, τα αποτελέσματα θα δημοσιευθούν σε ανώνυμη μορφή και οι συμμετέχοντες θα έχουν πρόσβαση σε αυτά, όπως προβλέπει το έντυπο συγκατάθεσης της έρευνας.
- Για την ανάγκη προγραμματισμού των συναντήσεων οι φοιτητές/ερευνητές θα χρησιμοποιήσουν στοιχεία επικοινωνίας των συμμετεχόντων, τα οποία όμως δεν θα συνδεθούν με τα δεδομένα που συλλέγονται κατά την ερευνητική διαδικασία και σε κάθε περίπτωση θα διαγράφονται με την ολοκλήρωση της συνάντησης.

7. **Ανάλυση Δεδομένων**:
   - Τα δεδομένα που θα προκύψουν από το ερωτηματολόγιο θα υποβληθούν σε ανάλυση για την αξιολόγηση των διαφορών στις διαγνωστικές αποφάσεις και στην αυτοπεποίθηση των γιατρών, πριν και μετά τη χρήση της ΤΝ. Παράλληλα, θα αναλυθούν τα προφίλ και οι συμπεριφορές των ιατρών κατά την αλληλεπίδρασή τους με τις τεχνηέντως λάθος υποδείξεις της ΤΝ.

Βιβλιογραφία

[1] J. R. Hampton, D. Adlam, and J. Hampton, 150 ECG Cases. Edinburgh: Elsevier, 2019.

[2] S. Günay, A. Öztürk, H. Özerol, Y. Yiğit, and A. K. Erenler, “Comparison of emergency medicine specialist, cardiologist, and chat-GPT in electrocardiography assessment,” The American Journal of Emergency Medicine, vol. 80, pp. 51–60, Jun. 2024. doi:10.1016/j.ajem.2024.03.017

[3] “The world’s most popular survey platform,” SurveyMonkey, https://www.surveymonkey.com/ (accessed Dec. 13, 2024).

Appendix G. Study Protocol (English Version)

Brief Description of the Study

Theoretical Background

Artificial Intelligence (AI) is reshaping the field of diagnostic practices, offering new tools that can enhance the quality and speed of medical decisions. Especially in cardiology, where accurate and immediate diagnosis of cardiovascular diseases is critical to treatment outcomes, the integration of advanced AI systems opens new horizons. Modern systems that utilize Large Language Models (LLMs) with Retrieval-Augmented Generation (RAG) technology allow cardiologists to access a vast volume of specialized medical knowledge in real time, offering diagnostic suggestions based on each patient's unique data.

LLMs, using RAG technology, function as "digital assistants" or AI co-pilots, providing suggestions based on individual and comparative patient data. This support facilitates the decision-making process for doctors, allowing them to identify potential diagnoses and reassess their initial evaluations. Thanks to the ability to integrate information from previous medical data, these systems help deliver analytical and accurate suggestions, strengthening clinical reasoning and physician confidence.

Despite their capabilities, these systems do not replace the knowledge and experience of the physician but rather function collaboratively, allowing doctors to maintain control over the diagnostic process. This interactive collaboration between humans and AI creates a new model in clinical practice, where AI support is utilized in combination with the judgment and experience of the physician, forming a human-machine interaction environment that fosters trust.

However, the application of AI in medicine remains an emerging field with many open challenges. Issues such as the objectivity of diagnostic suggestions, acceptance of these systems by doctors, their critical stance towards them, and the viability of their collaboration in real clinical settings require further investigation. The need for thorough evaluation of these systems is crucial, as it will help us understand how human-AI interaction can enhance accuracy and effectiveness in diagnosis, contributing to their smooth integration into daily medical practice and ensuring quality patient care.

Research Procedure

The research focuses on the interaction of cardiologists and internists with an Artificial Intelligence (AI) system that supports the diagnostic process. The aim is to evaluate the extent to which the integration of AI can improve the accuracy and confidence of doctors in diagnosing cardiology cases, as well as to assess the trust and critical stance of physicians toward AI. Participants will evaluate clinical cases with and without the support of AI, through the completion of a questionnaire. The process will last approximately 50–60 minutes in total. This study is conducted as part of two master's theses for the Interdepartmental Postgraduate Program in Translational Biomedical Engineering and Science.

The present study includes the following stages:

1. Selection and Description of Participants:

- The study targets cardiologists and internists. Recruitment of participating doctors is conducted through searches in academic, professional, and personal networks. Doctors who meet the participation criteria are informed personally about the purpose and procedure of the study. The briefing includes a detailed explanation of the study's aim, the nature of their interaction with the AI system, as well as the potential benefits and risks of their participation. At the beginning of each session, the participant states their specialty, work experience, and whether they completed their specialization in Greece. Additionally, they declare their level of familiarity with AI applications.

2. Presentation of Clinical Cases:

- The questionnaire participants are asked to complete consists of clinical cases that include electrocardiogram (ECG) data along with basic patient demographic information (age, gender, symptoms). These cases are drawn from a cardiologist training manual [1]. The cases have been converted into multiple-choice questions in a previous study [2]. For the purposes of the current study, an additional 10 questions from the manual were used, which were converted into multiple-choice questions with the assistance of a cardiologist. The purpose of adding these questions, as described in more detail below, is to observe and analyze physicians' attitudes and judgment toward questions where the AI suggestion is invalid.

3. Initial Diagnosis:

- Participants examine each case separately and record their initial diagnosis based on the ECG provided data and their own clinical experience.

4. Intervention and Suggestion by AI:

- After recording the initial diagnosis, the same case is presented again with a diagnostic suggestion from the AI. This suggestion has been generated by an LLM system. Doctors can re-evaluate their initial assessment using the AI suggestion as a guide or confirm their original decision.
- Among the suggestions provided by the AI model, some (~20% of the total) will be intentionally and selectively incorrect (without prior notification to the physician). With the help of these suggestions, the aim is to record and analyze the influence of AI on physicians of varying levels of expertise and familiarity with AI applications, as well as their behavior when faced with cases where the model's knowledge level does not correspond to that of the physician. This process is designed and implemented in such a way that it does not affect the cases where the physician interacts with the actual suggestions of the AI model.

5. Session Completion:

- Upon completion of the questionnaire by each doctor, the participant is informed about the presence of incorrect suggestions in order to understand the doctor's experience while responding to the questions and interacting with the (correct and incorrect) AI suggestions.

6. Data Collection and Recording:

- The full name of each participant will appear only on the consent form signature, which will not be digitized and will not be linked to the other data collected. At the same time, in order to make it possible for participants to withdraw their data from the study, they will be given a random code at the start of the process, which they will record themselves in their questionnaire. The connection between name and code will be known only to the participant, who, if they wish to withdraw their responses using the code, enables the researchers/students to identify their answers.
- Apart from the full name, no other personal information will be recorded. All data collected in the context of the study will remain completely anonymous. For the purposes of the study, the participants' specialty (cardiologist or internist), their professional experience (years since specialization), whether they completed their specialization in Greece, and their level of familiarity with AI applications will be recorded.
- All anonymous study data is collected through the questionnaire implemented on the SurveyMonkey service [3]. Upon completion of the data collection process, the anonymous responses will be deleted from the SurveyMonkey service and stored for analysis. According to the terms of use of SurveyMonkey, deletion of responses is permanent, access is lost immediately, and their complete deletion from the service's systems occurs within 60 days.
- Access to the data will be strictly limited to the researchers responsible for the study and will not be used for purposes beyond the current study. Regarding possible intellectual property rights of participants, the collected information does not include data that could identify individuals, thereby ensuring that there is no violation of their rights. Additionally, the results will be published in an anonymous form, and participants will have access to them, as stated in the study's consent form.
- For the purpose of scheduling meetings, the students/researchers will use the contact information of the participants, which, however, will not be linked to the data collected during the research process and will in any case be deleted upon completion of the meeting.

7. Data Analysis:

- The data obtained from the questionnaire will be analyzed to assess differences in diagnostic decisions and physician confidence before and after the use of AI. At the same time, the profiles and behaviors of doctors during their interaction with the deliberately incorrect AI suggestions will be analyzed.

Bibliography

[1] J. R. Hampton, D. Adlam, and J. Hampton, 150 ECG Cases. Edinburgh: Elsevier, 2019.

[2] S. Günay, A. Öztürk, H. Özerol, Y. Yiğit, and A. K. Erenler, "Comparison of emergency medicine specialist, cardiologist, and chat-GPT in electrocardiography assessment," The American Journal of Emergency Medicine, vol. 80, pp. 51–60, Jun. 2024. doi:10.1016/j.ajem.2024.03.017

[3] "The world's most popular survey platform," SurveyMonkey, https://www.surveymonkey.com/ (accessed Dec. 13, 2024).