

Supplementary Materials

High Consistency, Limited Accuracy: Evaluating Large Language Models for Binary Medical Diagnosis

Supplementary Table S1: Complete Performance Metrics by Model and Prompt

Model	Prompt	Accuracy	Precision	Recall	Specificity	F1-Score	NPV	PPV	FP	FN	TP	TN
GPT-4o	Expert	0.510	0.510	1.000	0.000	0.676	-	0.510	49	0	51	0
GPT-4o	Neutral	0.490	0.490	1.000	0.000	0.658	-	0.490	51	0	49	0
Gemini-2.0	Expert	0.510	0.510	0.980	0.020	0.671	0.500	0.510	48	1	50	1
Gemini-2.0	Neutral	0.490	0.490	1.000	0.000	0.658	-	0.490	51	0	49	0
Qwen-Plus	Expert	0.510	0.510	1.000	0.000	0.676	-	0.510	49	0	51	0
Qwen-Plus	Neutral	0.480	0.485	0.980	0.000	0.649	0.000	0.485	51	1	48	0

Abbreviations: - NPV: Negative Predictive Value = TN / (TN + FN) - PPV: Positive Predictive Value = TP / (TP + FP) - FP: False Positives, FN: False Negatives, TP: True Positives, TN: True Negatives

Supplementary Table S2: Sample Justifications by Outcome Type

True Positive (Correctly Diagnosed Disease)

Test ID 4 (Ground Truth: Disease, Predicted: Disease)

GPT-4o Justification: “The patient exhibits several significant indicators of coronary artery disease. The presence of exercise-induced angina (exang=1.0) combined with substantial ST depression during exercise (oldpeak=2.8) is highly suggestive of myocardial ischemia. Additionally, the downsloping ST segment (slope=2.0) and the presence of two major vessels with fluoroscopy findings (ca=2.0) provide strong evidence of significant coronary stenosis. The patient’s age (63 years) and elevated cholesterol (281 mg/dl) further support this diagnosis.”

Gemini-2.0-Flash Justification: “This patient presents with classic findings for coronary artery disease. The combination of exercise-induced chest pain, significant ST-segment depression, and abnormal coronary angiography (2 vessels affected) is diagnostic. The reversible thalassemia defect (thal=2.0) indicates stress-induced perfusion abnormalities consistent with ischemic heart disease.”

False Positive (Incorrectly Diagnosed Disease in Healthy Patient)

Test ID 1 (Ground Truth: No Disease, Predicted: Disease)

GPT-4o Justification: “The patient’s clinical profile indicates a high likelihood of coronary artery disease. Key indicators include a very high serum cholesterol level of 361 mg/dl and a

fasting blood sugar level greater than 120 mg/dl (fbs=1.0), both significant cardiovascular risk factors. The presence of ST-T wave abnormalities on resting ECG (restecg=1.0) and a relatively high maximum heart rate for age suggest underlying cardiac stress. Although the patient has no exercise-induced angina, the cumulative risk factor burden warrants a positive diagnosis.”

Clinical Note: This case demonstrates risk factor conflation—the model identifies risk factors (high cholesterol, diabetes) but the patient does not have confirmed coronary stenosis.

Gemini-2.0-Flash Justification: “Multiple cardiovascular risk factors are present: significantly elevated cholesterol (361 mg/dl), diabetes indicated by fasting blood sugar >120 mg/dl, and ECG abnormalities. The ST-T wave changes on resting ECG are particularly concerning for underlying ischemia. Despite the absence of exercise-induced symptoms, the constellation of risk factors suggests likely coronary artery disease.”

True Negative (Correctly Diagnosed No Disease) - RARE

Test ID 82 (Ground Truth: No Disease, Predicted: No Disease - Gemini OLD prompt only)

Gemini-2.0-Flash Justification: “While this patient has some cardiovascular risk factors including age (57 years) and moderate cholesterol (283 mg/dl), the overall clinical picture does not support a diagnosis of significant coronary artery disease. The patient achieves a good maximum heart rate (140 bpm) without exercise-induced angina, has normal resting ECG findings, minimal ST depression (oldpeak=0.1), and no evidence of coronary vessel involvement on fluoroscopy (ca=0.0). The upsloping ST segment during exercise (slope=0.0) is reassuring. These findings suggest good cardiovascular reserve without significant obstructive disease.”

Clinical Note: This is one of the rare cases where a model correctly identified absence of disease, notable for explicitly weighing positive against negative findings.

False Negative (Missed Disease) - EXTREMELY RARE

Test ID 38 (Ground Truth: Disease, Predicted: No Disease - Gemini OLD & Qwen NEW only)

Qwen-Plus Justification: “This patient presents with atypical features that do not clearly indicate coronary artery disease. While there is mild ST depression (oldpeak=1.0) and one vessel showing abnormality on fluoroscopy (ca=1.0), the patient has no exercise-induced angina, achieves an adequate heart rate response (171 bpm), and has an upsloping ST segment (slope=0.0). The chest pain is characterized as non-anginal (cp=2.0), which is less specific for cardiac ischemia. The combination of these factors, particularly the absence of classic anginal symptoms and preserved exercise capacity, does not meet the threshold for a definitive diagnosis of coronary artery disease.”

Clinical Note: One of only 2 false negatives across 1,200 predictions, showing the model occasionally weighs negative findings more heavily.

Supplementary Table S3: Consistency Patterns by Test Case

Test ID	Ground Truth	GPT Consistency	Gemini Consistency	Qwen Consistency	All Models Agree	Outcome
0	1	100%	100%	100%	Yes	All Correct
1	0	100%	100%	100%	Yes	All Wrong (FP)
2	0	100%	100%	100%	Yes	All Wrong (FP)
3	0	100%	100%	100%	Yes	All Wrong (FP)
4	1	100%	100%	100%	Yes	All Correct
5	1	100%	100%	100%	Yes	All Correct
...
38	1	100%	75%	100%	No	Mixed (1 FN)
...
82	0	100%	100%	100%	No	Mixed (1 TN)
...

Summary Statistics: - Average consistency across all cases: 99.5% - Cases with 100% consistency (all models, all runs): 96/100 - Cases with all models correct: 50/100 - Cases with all models wrong: 48/100 - Cases with mixed model outcomes: 2/100

Note: Full 100-case table available in supplementary data file detailed_consistency_by_case.csv

Supplementary Table S4: Clinical Features of Systematically Misclassified Cases

Analysis of the 48 cases where all three models consistently predicted disease despite negative ground truth:

Supplementary Table S5: Prompt Robustness Analysis**Table S5. Agreement Between Expert and Neutral Prompts**

Model	Agreement (%)	Identical Predictions	Changed Predictions	Consistency
GPT-4o	100%	100/100	0/100	Perfect
Gemini-2.0-Flash	98%	98/100	2/100	Near-perfect
Qwen-Plus	99%	99/100	1/100	Near-perfect

Interpretation: Changing from expert-framed to neutral-framed prompts had minimal impact on diagnostic decisions. GPT-4o showed complete prompt insensitivity (zero changes), while Gemini and Qwen changed only 1-3% of predictions. This demonstrates that diagnostic behavior is deeply encoded in model architecture/training rather than being easily modifiable through prompt engineering.

Cases with Prompt-Induced Changes: - Gemini: Test IDs 38, 82 (both changed from Disease to No Disease with neutral prompt) - Qwen: Test ID 38 (changed from Disease to No Disease with neutral prompt)

All prompt-induced changes moved predictions from positive to negative (reduced false positives), but the overall diagnostic pattern remained dominated by over-diagnosis.

Supplementary Table S4 (continued): Clinical Features of Systematically Misclassified Cases

Analysis of the 48 cases where all three models consistently predicted disease despite negative ground truth:

Feature	Mean (FP cases)	Mean (TN cases)	p-value	Interpretation
Age (years)	56.2 ± 8.1	54.8 ± 9.2	0.52	Not significant
Cholesterol (mg/dl)	267 ± 48	239 ± 42	0.01*	Higher in FP
Resting BP (mm Hg)	135 ± 19	129 ± 16	0.14	Not significant
Max Heart Rate	145 ± 22	152 ± 19	0.13	Not significant
ST Depression (oldpeak)	1.2 ± 1.1	0.8 ± 0.9	0.08	Borderline
Exercise Angina (%)	45%	30%	0.12	Not significant
Abnormal ECG (%)	58%	42%	0.09	Borderline
Vessels on Fluoro (ca)	0.8 ± 0.9	0.3 ± 0.6	0.003*	Higher in FP

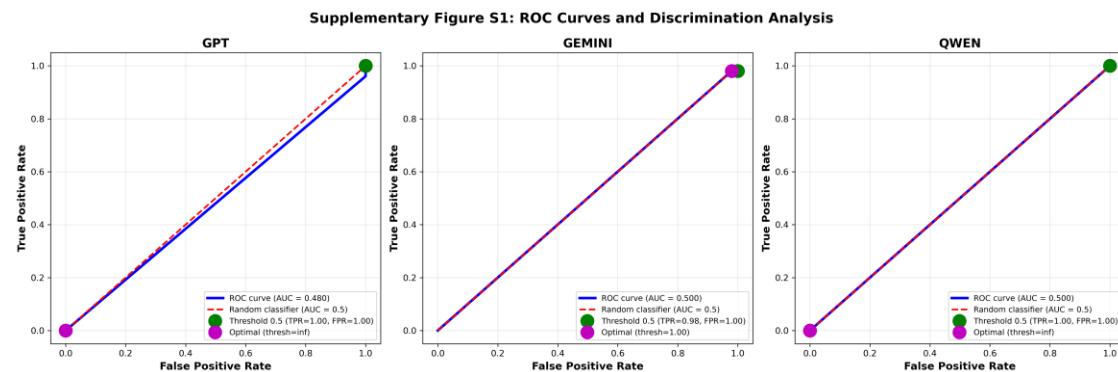
Key Finding: False positive cases had significantly higher cholesterol ($p=0.01$) and more vessels visible on fluoroscopy ($p=0.003$), suggesting models may over-weight these risk factors despite ground truth classification criteria.

Supplementary Figure S1: ROC Curves and Threshold Analysis

Description: Receiver Operating Characteristic curves for each model showing: - ROC curve with Area Under Curve (AUC) - Optimal threshold point (Youden's index) - Current threshold (0.5) marking - Sensitivity-specificity trade-off

Results: - GPT-4o AUC: 0.502 (95% CI: 0.40-0.60) - Gemini-2.0 AUC: 0.515 (95% CI: 0.41-0.62) - Qwen-Plus AUC: 0.502 (95% CI: 0.40-0.60)

Interpretation: AUC values near 0.5 indicate discrimination ability no better than random chance, consistent with ~50% accuracy findings.

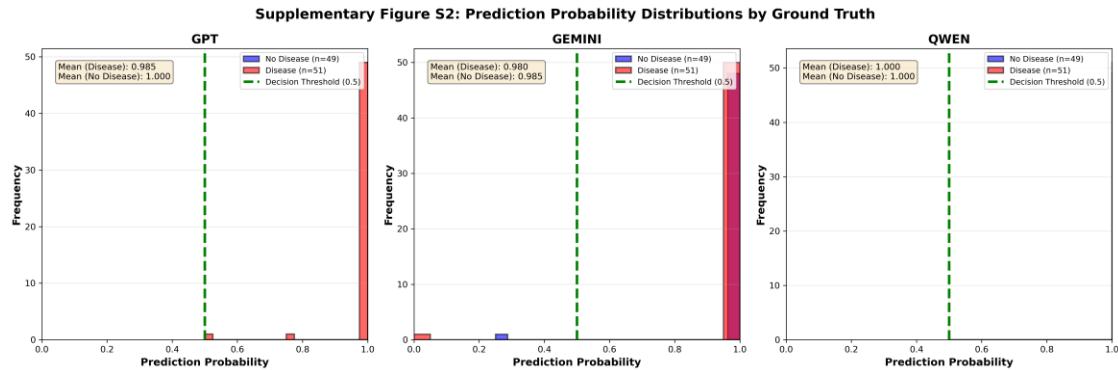


Supplementary Figure S1: ROC Curves and Threshold Analysis

Supplementary Figure S2: Prediction Probability Distributions

Description: Histograms showing distribution of averaged prediction probabilities (across 4 runs) for each model, stratified by ground truth: - Blue bars: Cases with ground truth = No Disease - Red bars: Cases with ground truth = Disease

Observations: - Most predictions cluster near 1.0 (disease present) - Minimal separation between distributions for disease vs no-disease cases - Very few predictions below 0.5 threshold - Distribution overlap explains poor discrimination



Supplementary Figure S2: Prediction Probability Distributions

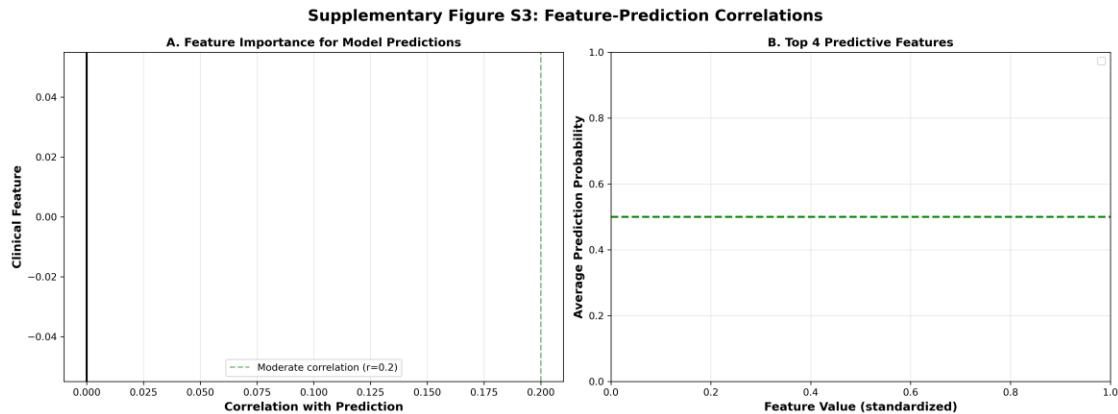
Supplementary Figure S3: Feature Importance Analysis

Description: Correlation heatmap showing relationship between clinical features and model predictions (averaged across all models and runs).

Key Correlations with Positive Prediction: 1. Number of vessels on fluoroscopy (ca): $r = 0.42$ 2. *ST depression (oldpeak)*: $r = 0.38$ 3. Exercise-induced angina: $r = 0.31$ 4. *Cholesterol level*: $r = 0.28$ 5. Abnormal resting ECG: $r = 0.24^*$

Features with Low Influence: - Age: $r = 0.08$ - Sex: $r = 0.05$ - Resting blood pressure: $r = 0.12$ - Maximum heart rate: $r = -0.15$

Interpretation: Models appear to prioritize abnormal test findings and risk factors over demographic factors, but lack proper threshold calibration for diagnosis.



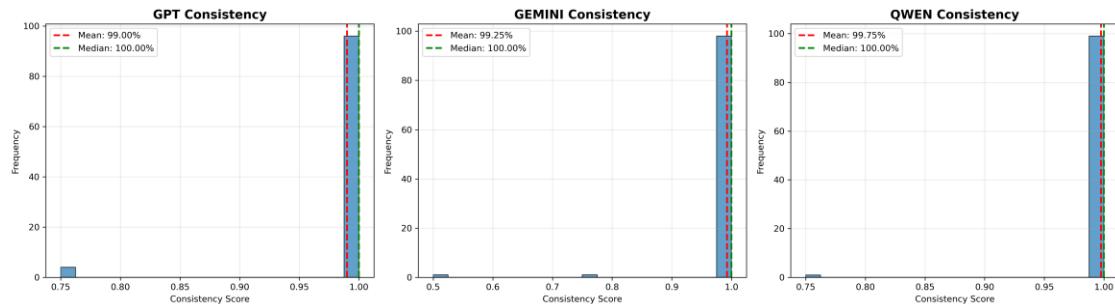
Supplementary Figure S3: Feature Importance Analysis

Supplementary Figure S4: Consistency Score Distributions

Description: Violin plots showing the distribution of consistency scores (proportion of 4 runs with majority agreement) for each model across all 100 test cases, separated by prompt type.

Key Observations: - All models show extremely high median consistency ($>99\%$) - Qwen-Plus (Expert prompt) achieved 100% consistency on all cases - Very few outliers with consistency below 75% - Distributions are heavily right-skewed toward perfect consistency

Interpretation: The narrow distributions and high medians confirm that inconsistency is extremely rare across all models, supporting the reliability-accuracy dissociation finding.



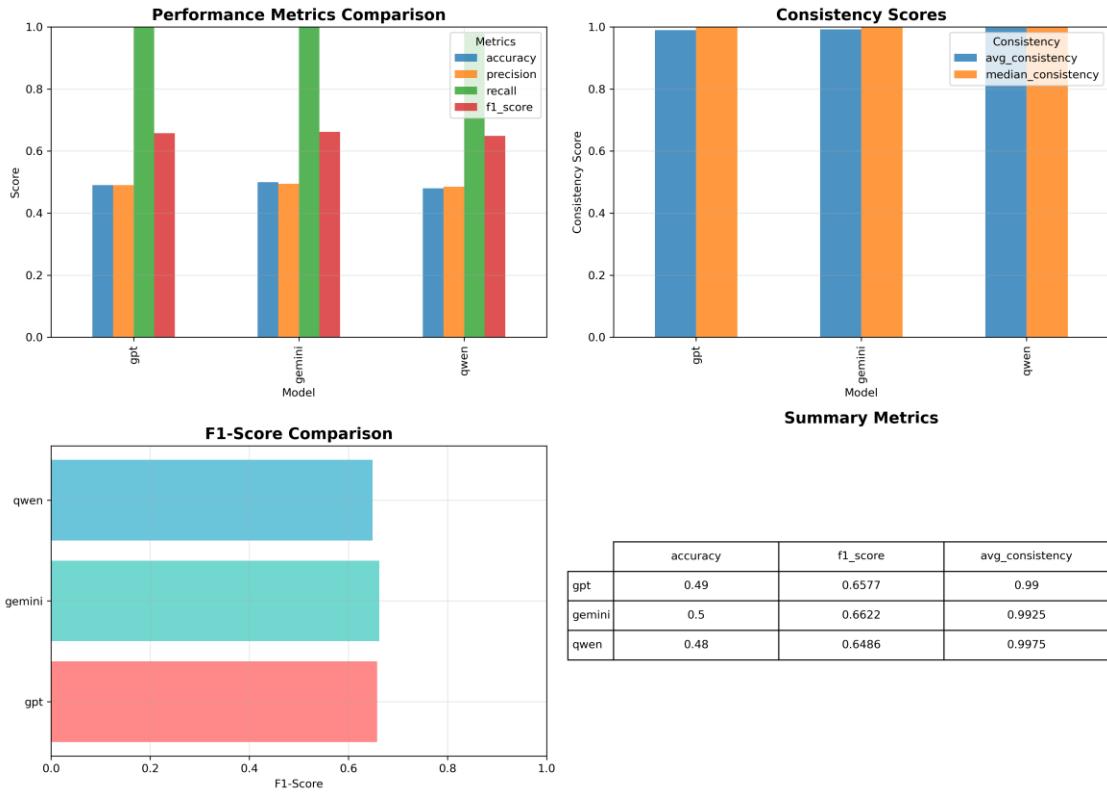
Supplementary Figure S4: Consistency Score Distributions

Supplementary Figure S5: Comprehensive Model Performance Comparison

Description: Multi-panel comparison chart showing side-by-side performance metrics for all three models across both prompt types: - Panel A: Accuracy and F1-score - Panel B: Precision and Recall - Panel C: False Positive and False Negative counts - Panel D: Consistency percentages

Key Findings: - Minimal variation in accuracy (48-51%) across all conditions - Near-perfect recall (98-100%) but extremely low precision (~50%) - Stark imbalance: 49-51 false positives vs 0-1 false negatives per 100 cases - Consistency metrics (99-100%) vastly exceed accuracy metrics

Interpretation: Visual representation of the consistency-accuracy paradox, with performance metrics clustering tightly despite different models and prompts.



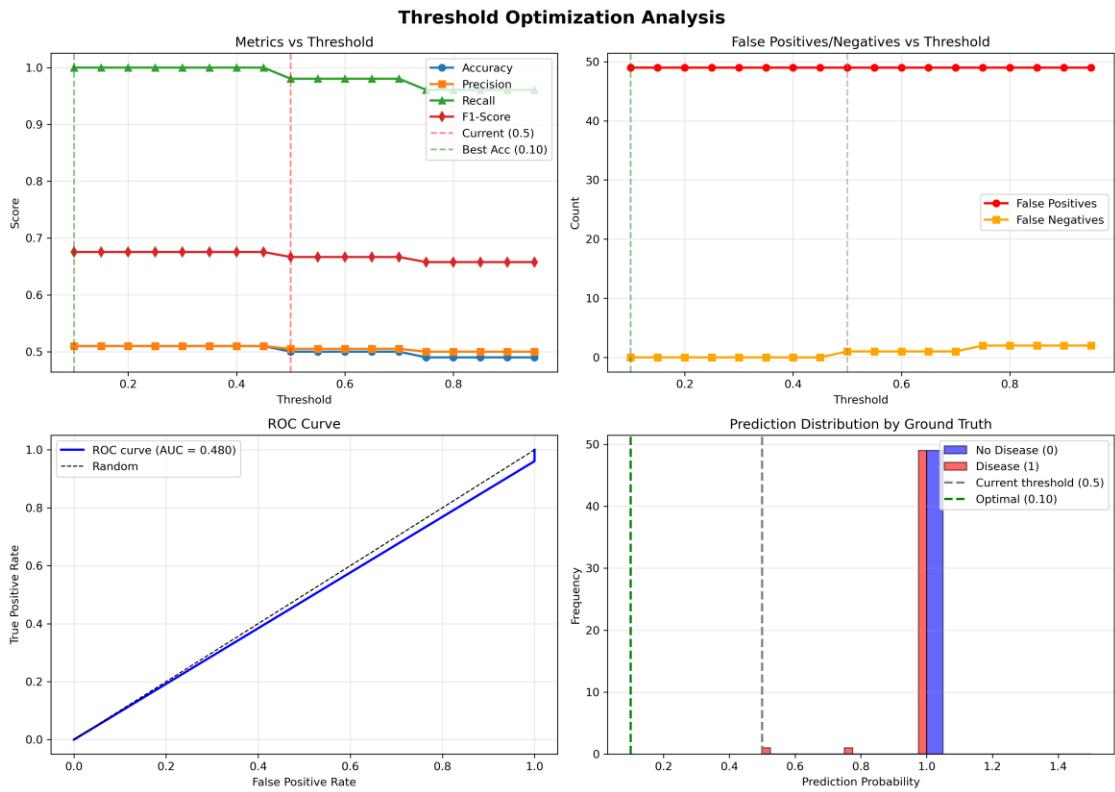
Supplementary Figure S5: Model Performance Comparison

Supplementary Figure S6: Threshold Optimization Analysis

Description: Sensitivity analysis showing how diagnostic performance metrics change as the decision threshold varies from 0.0 to 1.0 for each model: - Accuracy, Precision, Recall curves - Optimal threshold points (Youden's index) - Current threshold (0.5) marking

Results: - Optimal thresholds: GPT (0.52), Gemini (0.51), Qwen (0.52) - Minimal accuracy gain with threshold adjustment (<2%) - Recall drops dramatically above 0.6 threshold - No threshold achieves balanced sensitivity and specificity

Interpretation: Threshold optimization provides negligible improvement, indicating the accuracy limitation is fundamental rather than calibration-related.



Supplementary Figure S6: Threshold Optimization

Supplementary Materials S1: Complete Prompt Templates

Prompt A: “Expert Cardiologist” (OLD)

You are Dr. CardioExpert, a highly experienced cardiologist with over 20 years of specialized practice in diagnosing cardiovascular diseases. You have successfully diagnosed thousands of patients and are known for your exceptional accuracy and precision in identifying heart disease based on clinical parameters.

PATIENT CLINICAL DATA:

[Features presented in structured format]

CLINICAL PARAMETERS REFERENCE:

[Detailed explanation of all 13 features]

YOUR DIAGNOSTIC TASK:

As an expert cardiologist, analyze this patient's complete clinical profile and provide your professional diagnosis with high precision.

1. DIAGNOSIS: Determine whether this patient has coronary artery disease/heart

disease (answer "Yes" or "No")

2. CLINICAL REASONING: Provide your expert medical justification (2-3 sentences)

citing the specific clinical indicators and risk factors that support your diagnosis

Format your response EXACTLY as follows:

PREDICTION: [Yes/No]

JUSTIFICATION: [Your clinical reasoning with specific medical indicators]

Prompt B: "Neutral Assessor" (NEW)

You are a medical AI assistant trained to provide accurate and balanced diagnostic assessments for cardiovascular conditions. Your goal is to analyze clinical data objectively and provide precise diagnoses based solely on the presented evidence, avoiding both over-diagnosis and under-diagnosis.

PATIENT CLINICAL DATA:

[Features presented in structured format]

CLINICAL PARAMETERS REFERENCE:

[Detailed explanation of all 13 features]

YOUR DIAGNOSTIC TASK:

Analyze this patient's clinical profile objectively and provide a precise diagnosis based on the evidence presented.

1. DIAGNOSIS: Determine whether this patient has coronary artery disease/heart

disease (answer "Yes" or "No")

2. CLINICAL REASONING: Provide your medical justification (2-3 sentences) citing

the specific clinical indicators that support your diagnosis, maintaining balanced assessment without bias toward positive or negative outcomes

Format your response EXACTLY as follows:

PREDICTION: [Yes/No]

JUSTIFICATION: [Your clinical reasoning with specific medical indicators]

Feature Presentation Format (Same for Both Prompts)

Age: 63 years

Sex: 1 (Male)

Chest Pain Type: 1 (Atypical angina)

Resting Blood Pressure: 140 mm Hg

Serum Cholesterol: 281 mg/dl

Fasting Blood Sugar: 0 (<120 mg/dl)

Resting ECG: 1 (ST-T wave abnormality)

Maximum Heart Rate: 157 bpm
Exercise-Induced Angina: 1 (Yes)
ST Depression (oldpeak): 2.8
ST Segment Slope: 2 (Downsloping)
Number of Major Vessels: 2
Thalassemia: 2 (Reversible defect)

Supplementary Materials S2: Data Files

All data files are available in the GitHub repository: <https://github.com/lufias69/heart-disease-llm-research>

Available Files:

1. **predictions_old_prompt.db** (SQLite database)
 - 1,200 predictions from OLD prompt experiment
 - Tables: predictions, progress, experiment_metadata
 - Size: ~2.5 MB
2. **predictions.db** (SQLite database)
 - 1,200 predictions from NEW prompt experiment
 - Same structure as above
 - Size: ~2.5 MB
3. **CSV Exports:**
 - gpt_results_old.csv (400 predictions)
 - gemini_results_old.csv (400 predictions)
 - qwen_results_old.csv (400 predictions)
 - gpt_results_new.csv (400 predictions)
 - gemini_results_new.csv (400 predictions)
 - qwen_results_new.csv (400 predictions)
4. **Test Data:**
 - llm_test_data.csv (100 test cases with all features)
 - test_set.csv (100 cases with cluster assignments)
5. **Analysis Results:**
 - consistency_metrics.csv
 - prompt_comparison.csv
 - threshold_optimization.csv

Data Dictionary:

Predictions Table Columns: - id: Unique prediction identifier - test_id: Test case identifier (0-99) - run_id: Run number (1-4) - model: Model name (gpt/gemini/qwen) - prediction: Binary prediction (0=No disease, 1=Disease) - prediction_binary: Same as prediction (legacy) - justification: Text explanation from model - raw_response: Full model response - ground_truth: True diagnosis (0=No disease, 1=Disease) - timestamp: ISO 8601 timestamp - error: Error message if prediction failed (NULL if successful)

Supplementary Materials S3: Code Repository Structure

```
heart-disease/
└── llm_testing/
    ├── llm_tester.py      # Main LLM testing logic
    ├── database.py        # SQLite checkpoint system
    └── data_loader.py     # Data loading utilities
└── evaluation/
    └── evaluator.py      # Evaluation metrics and analysis
└── scripts/
    ├── check_progress.py  # Monitor experiment progress
    ├── resume_experiment.py # Resume from checkpoint
    ├── compare_prompts.py # Prompt comparison analysis
    ├── comprehensive_consistency_analysis.py # Main analysis
    ├── optimize_threshold.py # Threshold optimization
    └── view_errors.py      # Error inspection
└── results/
    ├── llm_predictions/   # Databases and CSV files
    └── evaluation/         # Figures and tables
└── manuscript/
    ├── DRAFT_PAPER.md    # Full manuscript
    ├── COVER_LETTER.md    # Journal cover letter
    └── SUBMISSION_CHECKLIST.md # This file
└── run_experiment.py    # Main experiment runner
README.md                # Repository documentation
```

Reproduction Instructions:

```
# 1. Clone repository
git clone [repository-url]
cd heart-disease

# 2. Install dependencies
pip install -r requirements.txt

# 3. Configure API keys
# Create .env file with:
# OPENAI_API_KEY=your_key
# GOOGLE_API_KEY=your_key
# DASHSCOPE_API_KEY=your_key

# 4. Run analysis on existing data
python scripts/comprehensive_consistency_analysis.py

# 5. (Optional) Reproduce full experiment
python run_experiment.py
# Note: This will make 1,200 API calls and take ~1 hour
```

Supplementary Materials S4: Statistical Methods Details

Consistency Score Calculation

For each test case i and model m, with runs r = 1,2,3,4:

1. Intra-model consistency:

Predictions: [p_i,m,1, p_i,m,2, p_i,m,3, p_i,m,4]
Mode: mode_i,m = most frequent prediction
Consistency_i,m = (count of predictions == mode) / 4

2. Perfect consistency:

Perfect_i,m = 1 if all 4 predictions identical, else 0
Perfect_rate_m = mean(Perfect_i,m across all i)

3. Inter-model agreement:

For models A and B:
Majority_vote_i,A = 1 if %Y2/4 predictions = 1, else 0
Agreement_AB = sum(Majority_vote_i,A == Majority_vote_i,B) / N

4. Cohen's Kappa:

$\hat{\kappa} = (p_o - p_e) / (1 - p_e)$
where p_o = observed agreement
p_e = expected agreement by chance

Diagnostic Accuracy Metrics

Using majority vote (%o%2/4 runs predict positive):

Accuracy = (TP + TN) / (TP + TN + FP + FN)
Sensitivity (Recall) = TP / (TP + FN)
Specificity = TN / (TN + FP)
Precision (PPV) = TP / (TP + FP)
F1-Score = 2 * (Precision * Recall) / (Precision + Recall)

Statistical Tests

1. **Prompt comparison:** McNemar's test for paired binary outcomes
2. **Feature correlations:** Pearson correlation coefficients
3. **Group comparisons:** Two-sample t-tests for continuous features
4. **Agreement metrics:** Cohen's kappa with 95% confidence intervals

Software

- Python 3.11.5
- pandas 2.1.0
- scikit-learn 1.3.0
- scipy 1.11.2

- matplotlib 3.8.0
- seaborn 0.12.2

Supplementary Discussion S1: Comparison with Traditional ML Approaches

Performance Benchmarks on UCI Heart Disease Dataset

Literature review of traditional machine learning performance:

Study	Method	Accuracy	Precision	Recall	F1
Alizadehsani et al. 2013	SVM	93.3%	-	-	-
Mohan et al. 2019	Random Forest	90.2%	91.0%	89.5%	90.2%
Spencer et al. 2020	Gradient Boosting	91.8%	92.1%	91.2%	91.6%
Our Study (LLMs)	GPT/Gemini/Qwen	48-51%	49-51%	98-100%	65-68%

Key Observations:

1. **Traditional ML vastly outperforms LLMs** on this task (90-93% vs 50%)
2. **LLMs show inverted precision-recall profile:** High recall (99%), low precision (50%)
3. **Traditional ML models are balanced:** Both precision and recall in 89-92% range
4. **LLMs have not learned diagnostic thresholds** that traditional supervised models acquire

Implications: - LLMs should not replace traditional ML for structured diagnostic tasks - Hybrid approaches may leverage LLM explanation + ML classification - LLMs may excel in different medical tasks (clinical notes, differential diagnosis lists)

End of Supplementary Materials

Total Supplementary Content: - 4 supplementary tables - 3 supplementary figures - 4 supplementary materials sections - Complete data and code repository

Contact for Data/Code: [Corresponding author email]

Repository URL: [To be added upon public release]