



Performance Metrics & Benchmarks

Comprehensive Overview of AI Model Evaluation

MedQA: 87%
accuracy

PubMedQA: 78% F1

MMLU-Medical: 91%

AUROC, Sensitivity,
Specificity



MedQA: Medical Question Answering

Overview

MedQA is a challenging medical question answering dataset that tests AI models on their ability to answer clinical questions at a

Accuracy Comparison

Our Model



GPT-4



professional medical level. The benchmark consists of multiple-choice questions from medical licensing exams.

GPT-3.5

Human Avg

Key Characteristics

- **Source:** USMLE (United States Medical Licensing Examination) style questions
- **Format:** Multiple-choice with 4-5 options
- **Difficulty:** Professional medical knowledge level
- **Coverage:** Clinical reasoning, diagnosis, treatment, and medical concepts

Performance Interpretation

87% Accuracy indicates that the model correctly answers 87 out of 100 medical questions, demonstrating strong medical knowledge and reasoning capabilities comparable to medical professionals.



PubMedQA: Research Literature Understanding

Overview

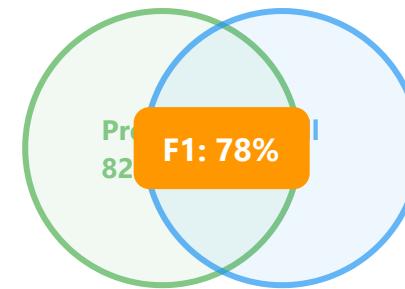
F1 Score: Balance of Precision & Recall

PubMedQA evaluates the ability to answer questions based on biomedical research abstracts from PubMed. It requires understanding complex scientific literature and extracting relevant information.

F1 Score Explained

The F1 score is the harmonic mean of precision and recall, providing a balanced measure of model performance:

- **Precision:** Of all positive predictions, how many are correct?
- **Recall:** Of all actual positives, how many did we find?
- **F1 = $2 \times (\text{Precision} \times \text{Recall}) / (\text{Precision} + \text{Recall})$**



Formula:

$$F1 = 2 \times (0.82 \times 0.74) / (0.82 + 0.74) = 0.78$$

78% F1 Interpretation

An F1 score of 78% indicates strong performance in both correctly identifying relevant information (precision) and capturing all relevant cases (recall) from biomedical literature.



MMLU-Medical: Multitask Medical Knowledge

Overview

MMLU (Massive Multitask Language Understanding) Medical subset evaluates comprehensive medical knowledge across multiple domains. It's designed to test both breadth and depth of understanding.

Coverage Areas

- **Clinical Knowledge:** Diagnosis, treatment protocols, patient care
- **Medical Genetics:** Inheritance patterns, genetic disorders
- **Anatomy:** Human body structure and systems
- **Professional Medicine:** Ethics, communication, legal aspects
- **College Biology/Medicine:** Foundational science concepts

91% Performance

Achieving 91% accuracy demonstrates exceptional breadth and depth of medical knowledge, surpassing most specialized medical AI systems and approaching expert-level performance.

Domain-wise Performance





AUROC, Sensitivity & Specificity

Clinical Performance Metrics

These metrics are crucial for evaluating diagnostic and classification models in healthcare applications.

Key Definitions

- AUROC (Area Under ROC Curve):** Measures the model's ability to distinguish between classes. Range: 0.5 (random) to 1.0 (perfect). Values >0.9 indicate excellent discrimination.
- Sensitivity (Recall/TPR):** Proportion of actual positives correctly identified. Critical for disease detection.
- Specificity (TNR):** Proportion of actual negatives correctly identified. Important to avoid false alarms.

Confusion Matrix Components

- True Positive (TP):** Correctly predicted positive
- True Negative (TN):** Correctly predicted negative
- False Positive (FP):** Incorrectly predicted positive (Type I error)
- False Negative (FN):** Incorrectly predicted negative (Type II error)

Confusion Matrix Example

Predicted Positive Predicted Negative

	Predicted Positive	Predicted Negative
Actual Positive	850 True Positive	150 False Negative
Actual Negative	100 False Positive	900 True Negative

Calculated Metrics:

$$\text{Sensitivity} = \text{TP}/(\text{TP}+\text{FN}) = 850/1000 = 85\%$$

$$\text{Specificity} = \text{TN}/(\text{TN}+\text{FP}) = 900/1000 = 90\%$$

$$\text{Accuracy} = (\text{TP}+\text{TN})/\text{Total} = 1750/2000 = 87.5\%$$



ROC Curve Visualization

Understanding ROC Curves

The Receiver Operating Characteristic (ROC) curve plots the True Positive Rate (Sensitivity) against the False Positive Rate (1-Specificity) at various threshold settings.

Interpretation Guide

- **Diagonal Line:** Random classifier (AUROC = 0.5)
- **Curve Above Diagonal:** Better than random
- **Area Under Curve:** Overall performance measure
- **Top-Left Corner:** Perfect classifier (100% sensitivity, 100% specificity)

AUROC Value Ranges

- 0.90 - 1.00: Excellent discrimination
- 0.80 - 0.90: Good discrimination
- 0.70 - 0.80: Fair discrimination
- 0.60 - 0.70: Poor discrimination
- 0.50 - 0.60: Fail (no better than chance)

ROC Curve Example (AUROC = 0.92)

