LIME for Clinical Text

Clinical Note Input

"Patient presents with **severe chest pain** radiating to left arm. History of **hypertension** and **diabetes**. ECG shows **ST elevation**. Recommend immediate cardiac intervention."

Step 1: Generate Perturbations

$$\xi = \text{argmin } L(f, g, \pi\_x) + \Omega(g)$$
Local fidelity + Complexity

✗ "severe chest pain"	✓ "severe chest pain"	✓ "severe chest pain"
✓ "hypertension"	✗ "hypertension, diabetes"	✓ "hypertension"
✓ "ST elevation"	✓ "ST elevation"	✗ "ST elevation"

P = 0.65   P = 0.78   P = 0.92

Generate N perturbed samples by randomly removing words

Step 2: Get Model Predictions   **Step 3: Fit Linear Model**

Black Box Model

Local Linear Approximation

| "chest pain": 0.35 | "ST elevation": 0.28 | "hypertension": 0.15 | Other: 0.22 |

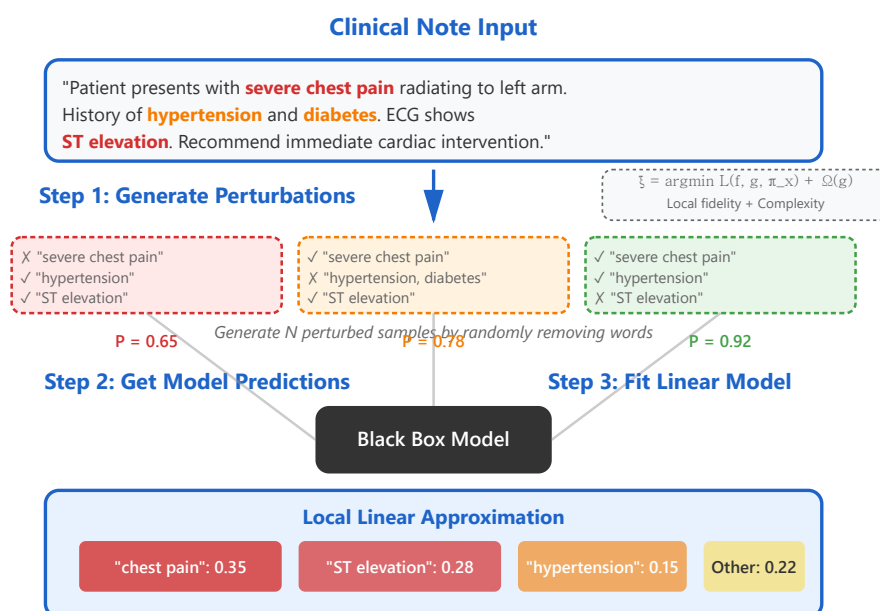## 🔍 Local Approximation

LIME explains individual predictions by approximating the black box locally

🔀 Text Perturbation

Randomly removes words to create variations of the original text

📊 Feature Importance

Linear model weights reveal word importance for the prediction

🏥 Clinical Application

Identifies key clinical terms driving diagnostic decisions