Reflection (5 points)

Most Challenging Part

The **preprocessing and feature engineering stage** posed the greatest challenge due to:

- **Data Heterogeneity**: Mixing structured and unstructured data (e.g., diagnosis codes + discharge notes) demands tailored NLP techniques.
- **Bias Detection**: Ensuring fairness across subpopulations added complexity and required domain-specific metrics.

Improvements with More Time/Resources

- **Enhanced Feature Enrichment**: Integrate external data sources (e.g., SDOH indices, wearable health data) to boost predictive insight.
- **Explainability Layer**: Incorporate SHAP or LIME to clarify model predictions for clinicians.
- **Continuous Learning**: Deploy MLOps tools to allow model retraining based on feedback loops and new patient data.

Workflow Diagram (5 points)

Here's a flowchart-style breakdown inspired by CRISP-DM and your use of NLP + model tuning practices:

Business Understanding

↓

Data Understanding

↓

Data Preparation

- Cleaning

- Normalization

- Feature Engineering

↓

Modelling

- Model Selection

- Training & Evaluation

- Confusion Matrix

↓

Evaluation

- Bias Auditing

- Precision/Recall

- SHAP/LIME (Explainability)

↓

Deployment & Monitoring

- API Integration

- HIPAA Compliance

- Feedback Loop