Approach

- Performed **data cleaning, outlier handling, and feature encoding** on the training set only to prevent data leakage.
- Built **reusable feature engineering functions** to ensure modularity and repeatability across experiments.
- Created **history based features**:
    - *One level history:* categorical columns target encoded by mean readmission rate per category.
    - *Two level history:* interactions of two categorical variables target encoded by mean readmission rate per pair.

- These features captured prior readmission tendencies at patient and encounter levels, improving model signal while avoiding leakage.
- Addressed **class imbalance** using class weights in all classifiers.
- Trained **Logistic Regression, Decision Tree, Random Forest, LightGBM, and XGBoost**, evaluated with ROC AUC, PR AUC, F1 score, and Brier score.
- Generated a **threshold based performance table** to show how precision, recall, F1, and flagged patient % change with different cutoffs.
- Thresholds were optimized on the validation set to support **collaborative calibration** with clinicians and administrators, balancing workload and risk.
- Prioritized **minimizing false negatives**, as missing high risk patients poses greater clinical risk.
- Used **coefficient analysis** to interpret influential predictors in the final Logistic Regression model.
- For **NER**, developed two extraction methods, rule based **spaCy NER** and generative **LLM (Phi 3 mini)** and compared their quality, and correctness.

Key Results

- **Final Model:** Logistic Regression gave best balance of interpretability and calibration (ROC AUC = 0.62, F1 = 0.55).
- **Top predictors:** prior admissions and diagnosis x medication complexity, aligning with medical intuition.
- **NER results:** LLM captured richer contextual details, spaCy offered more precise but narrower outputs, showing a trade off between coverage and reliability.

With more Time and Data

- Model can flag 60–65 % of high risk readmissions with explainable reasoning.
- Extracted entities enable **automated structured data capture** from clinical text.
- Combining **probability and age** supports **personalized risk based follow up**.
    - For example, a patient aged >70 years with >=80 % predicted readmission risk could trigger a more intensive follow up plan than younger, lower risk individuals.
- Future work: add cross validation, richer clinical features, fine tune domain LLMs, improve calibration, and apply cost based thresholding to align with clinical priorities.
- Create a **gold standard labeled subset** of discharge notes to compute metrics.
- **Hybrid modeling:** Combine both methods, use spaCy rules for anchor detection and feed these spans as structured hints to the LLM to balance accuracy and breadth.