MEDICAL IMAGE CAPTIONING FOR CHEST X-RAYS

Maria Linhares, 113534 Miguel Pinto, 107449

Problem & Motivation

Problem: Manual radiology report writing can be time-consuming and requires expertise
Opportunity: AI can assist radiologists by generating preliminary captions
Impact:

- Reduce radiologist workload
- Faster initial assessments
- Standardize reporting language
- Educational tool for medical students

Dataset Overview

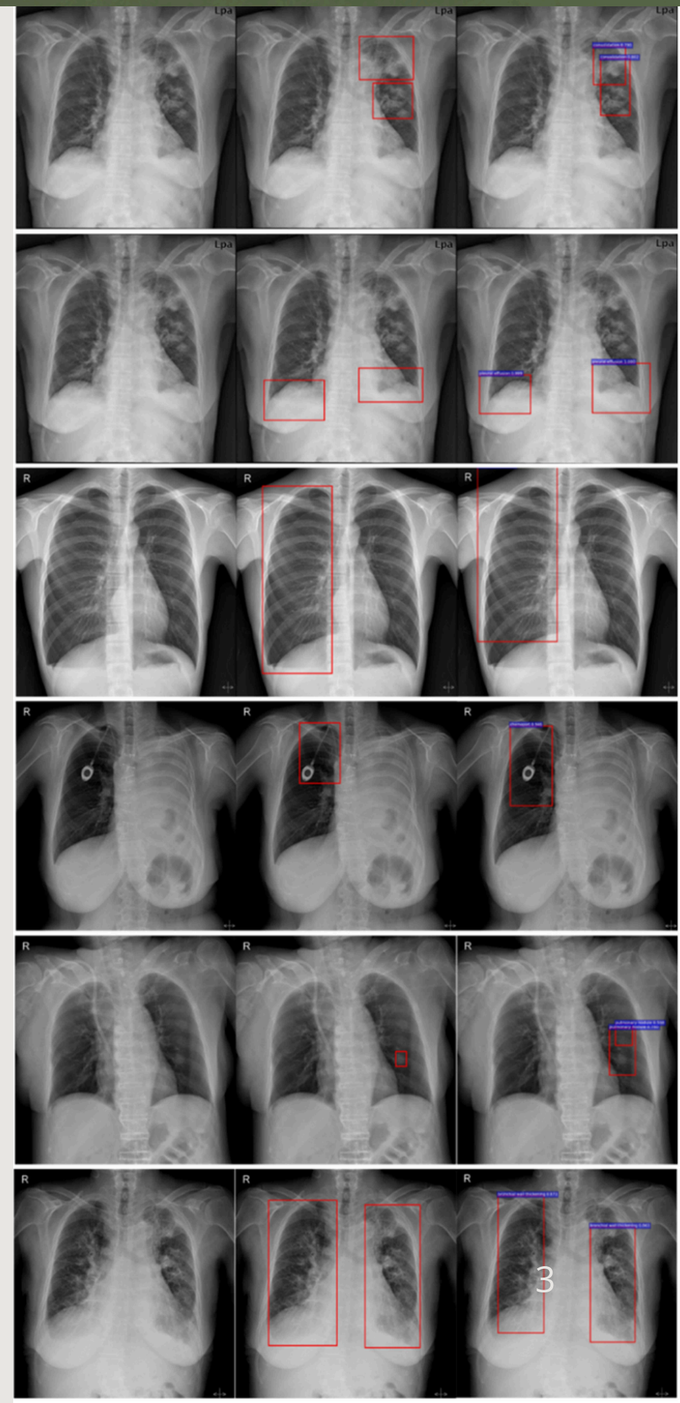
Indiana University Chest X-Ray Dataset

Statistics:

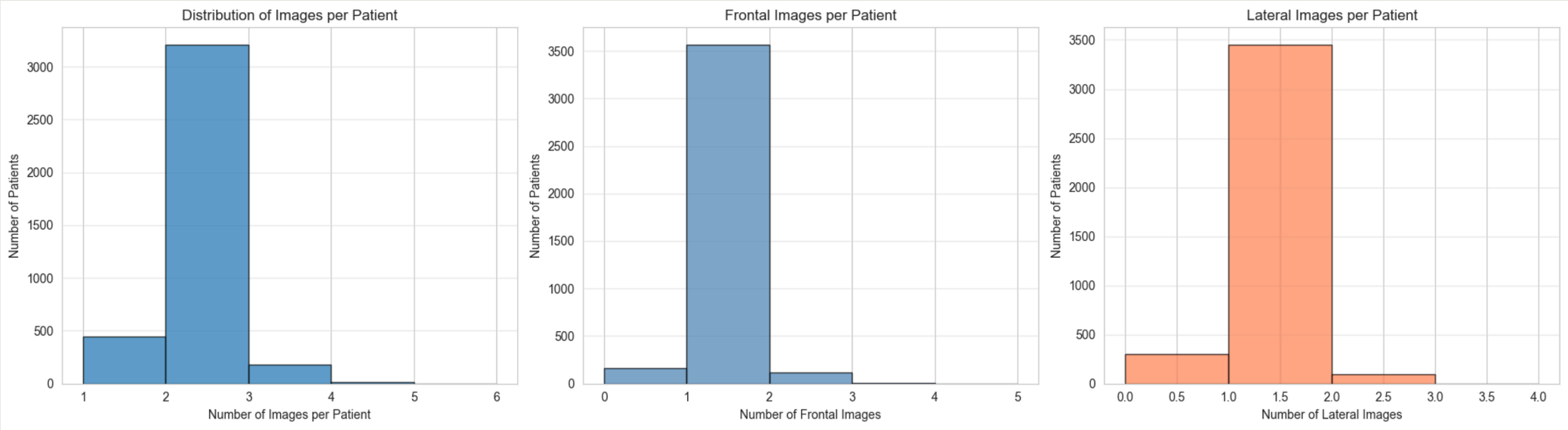
- 7,466 chest X-ray images
- 3,851 unique patients
- Frontal and lateral positions
- Associated radiology reports with “Findings” and “Impression” sections

Key Design Decision:

- Used **Impression section** as caption target (shorter, less censored)
- Filtered report with >30% anonymizations tokens (XXXX)



EDA - Image Analysis

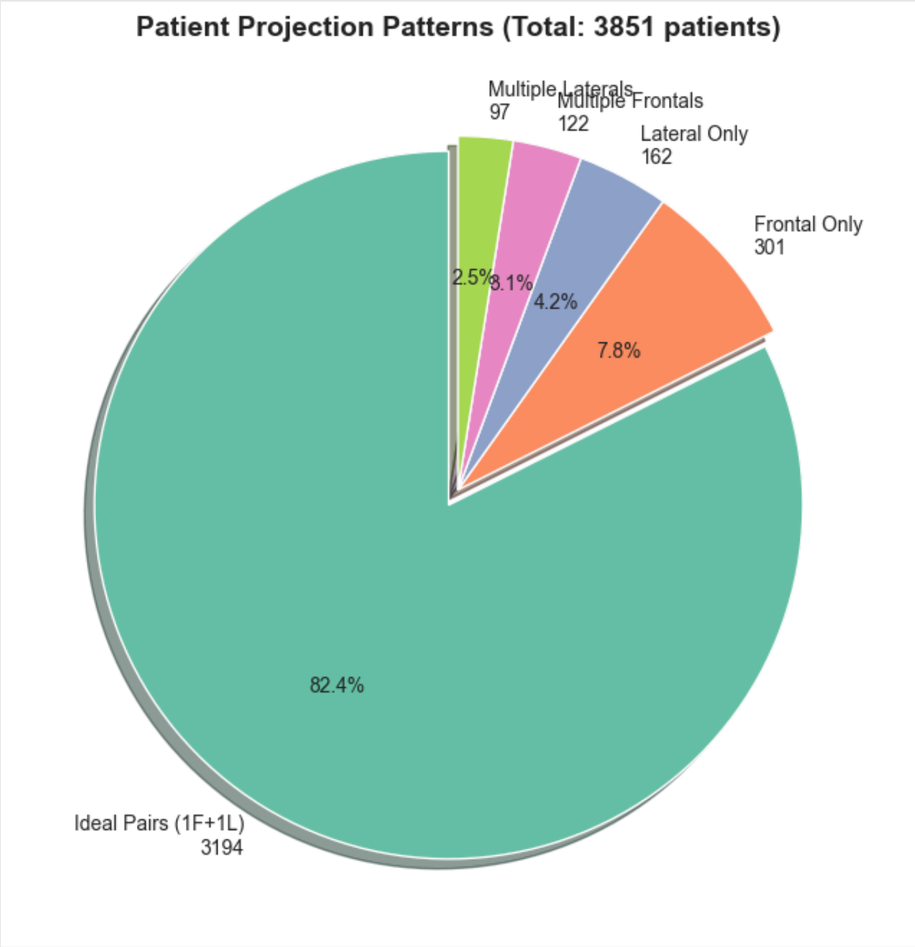


- Most have only 2 images (1 frontal + 1 lateral)
- Multiple images per patient per projection are common

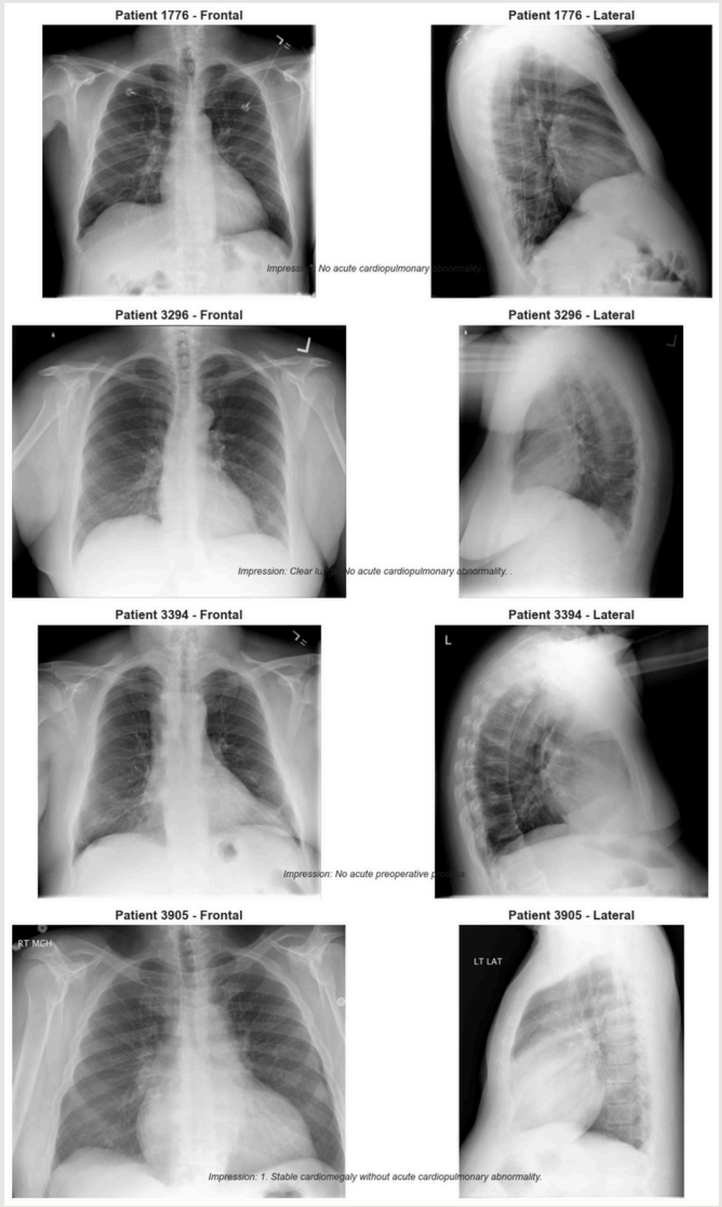
EDA - Image Analysis

1st Decision: Use 1st frontal image from patients
→ 3689 Patients
→ 1st frontal image per patient

2nd Decision: Use "Ideal Pairs" (1F+1L) patients
→ 3194 Patients
→ 1st frontal image per patient



EDA - Image Analysis



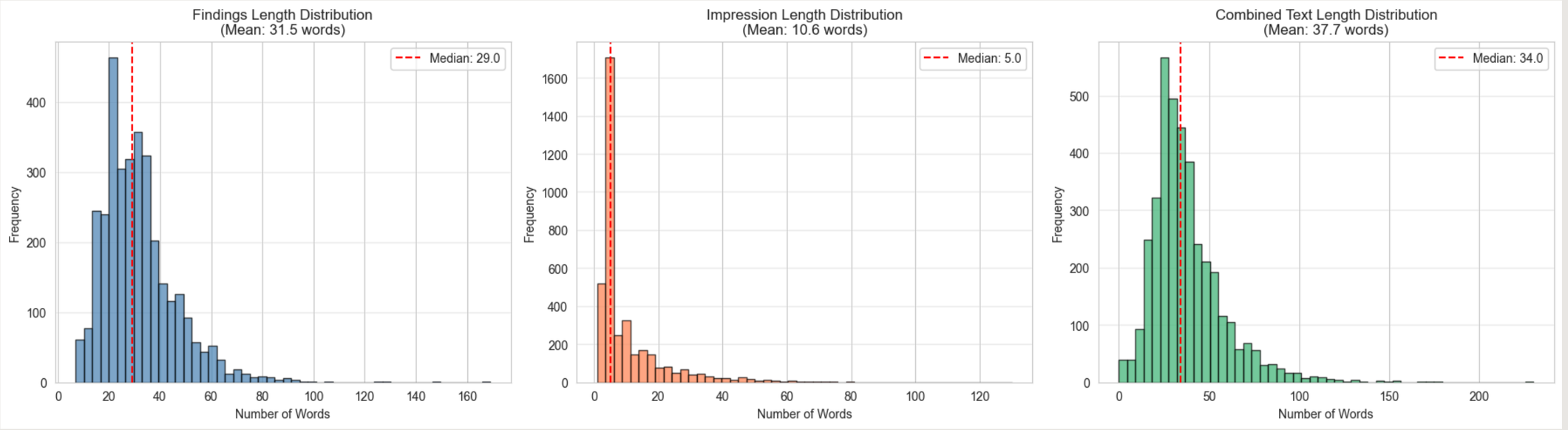
EDA - Text Analysis

- **Findings**: Detailed radiological observations
- **Impression**: Concise clinical conclusion

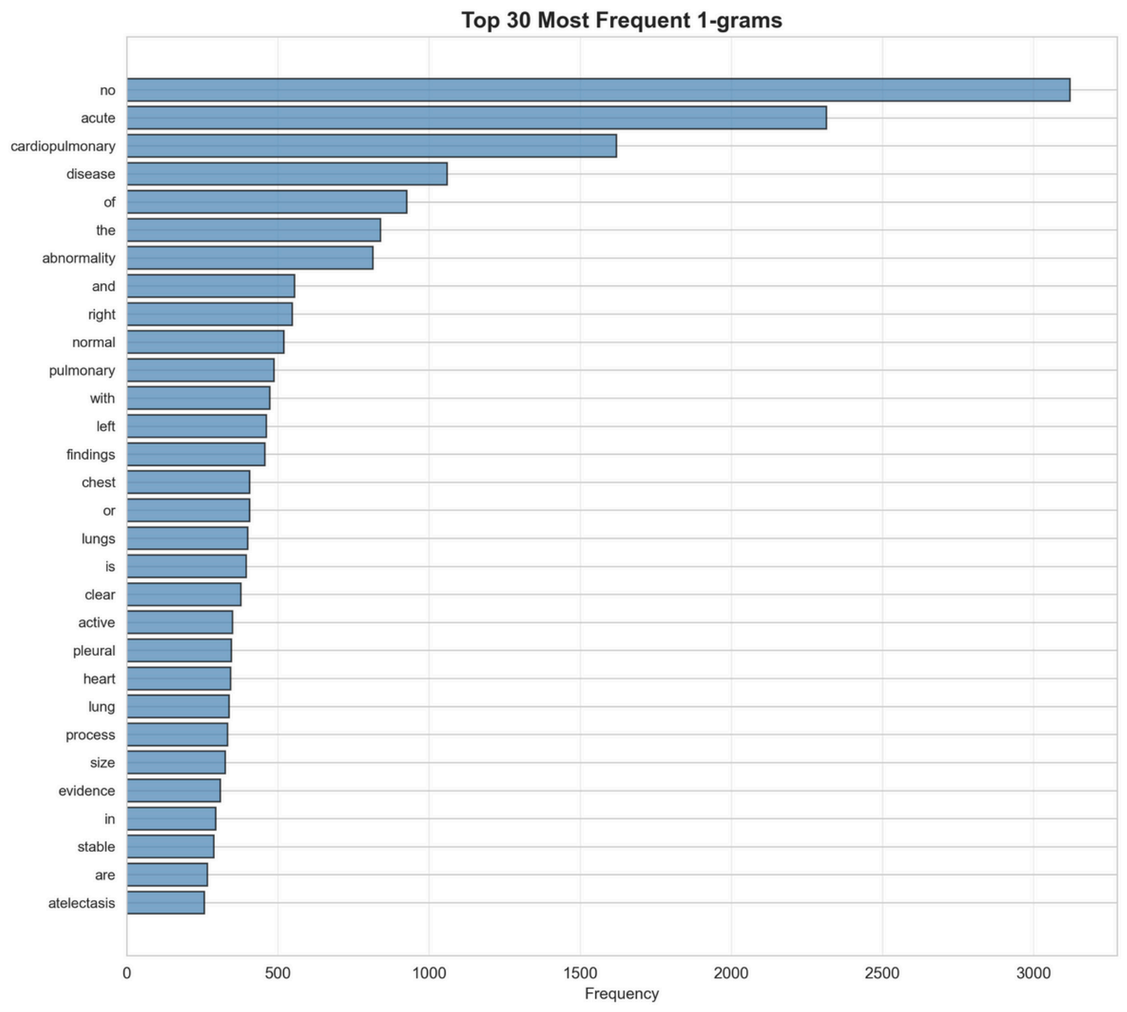
	Field	Missing	Missing %
0	findings	514	13.347183
1	impression	31	0.804986

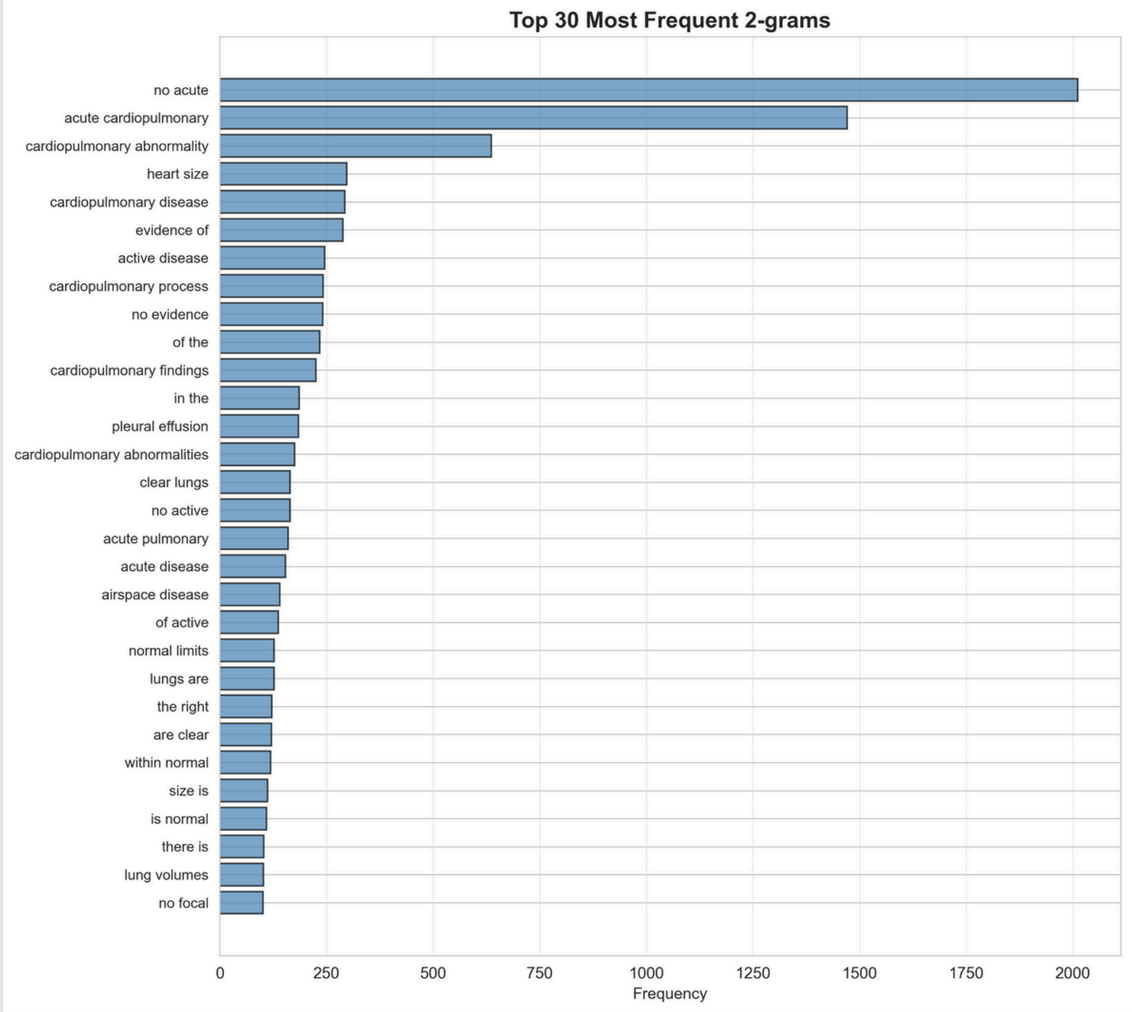
Caption Length (Impression):

- Mean: ~10,6 words
- Median: 5,0 words
- Max length cap: 100 words

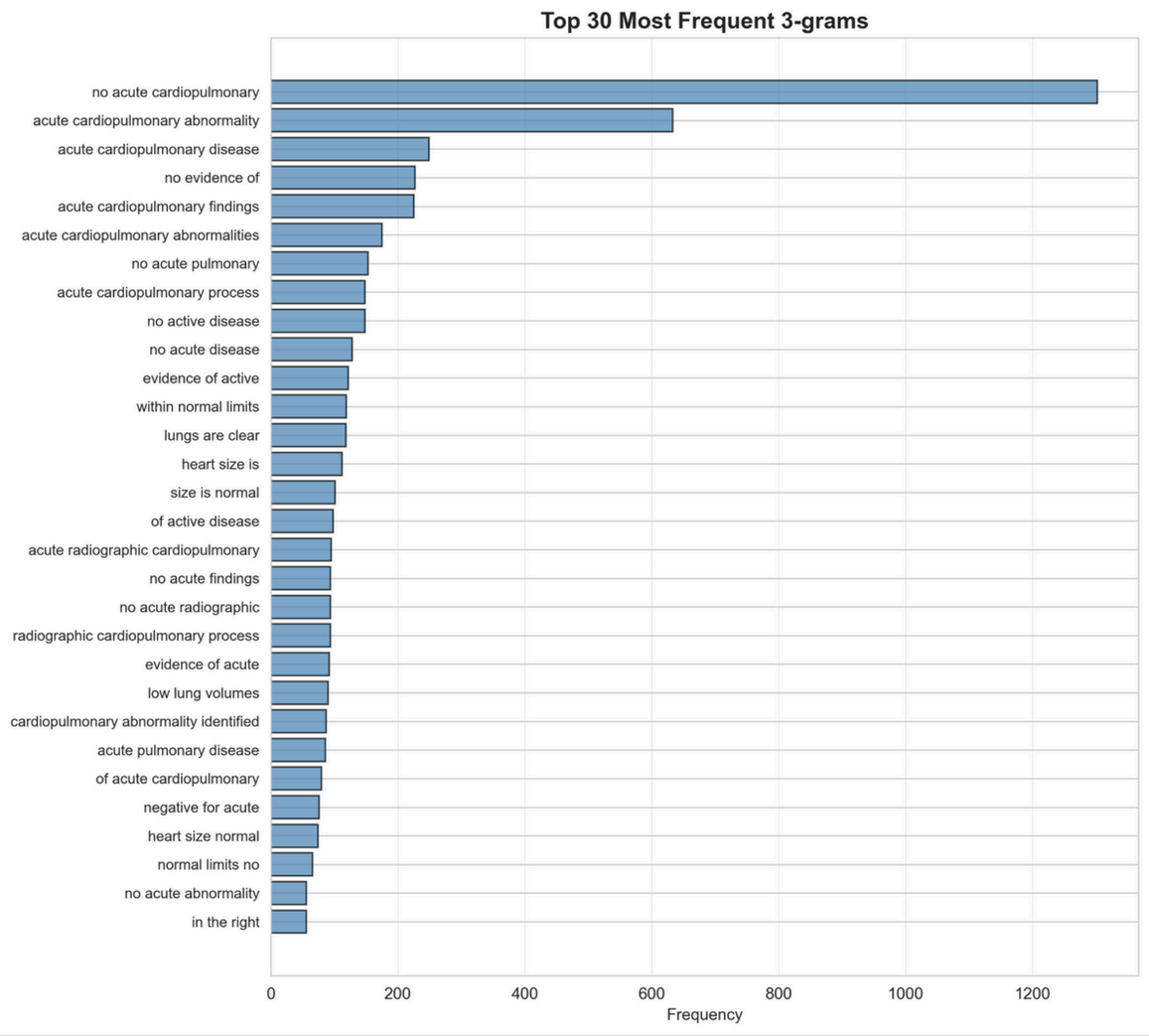


EDA - Text Analysis





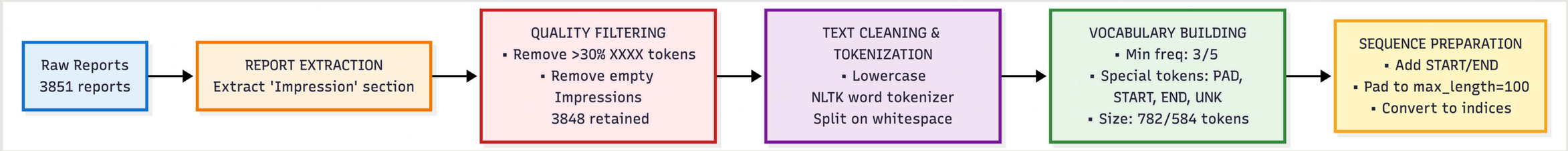
EDA - Text Analysis

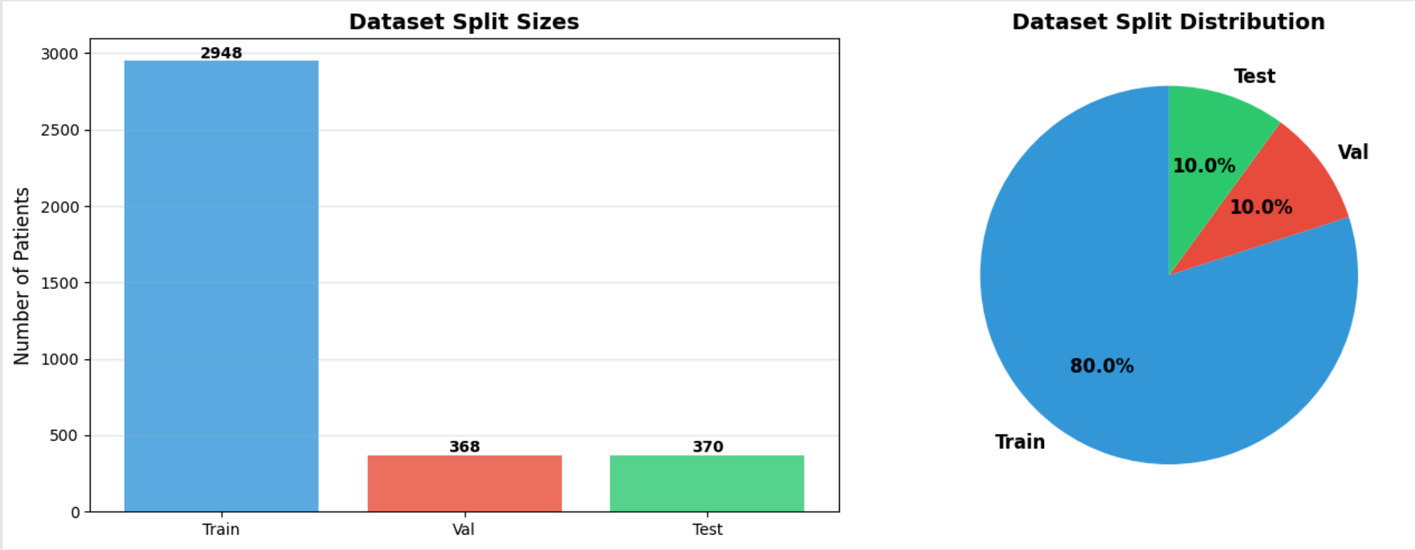


Vocabulary Statistics:

- Raw vocabulary: ~2000 unique tokens
- After min_freq=5: 514 tokens

Preprocessing - Pipeline & Data Split



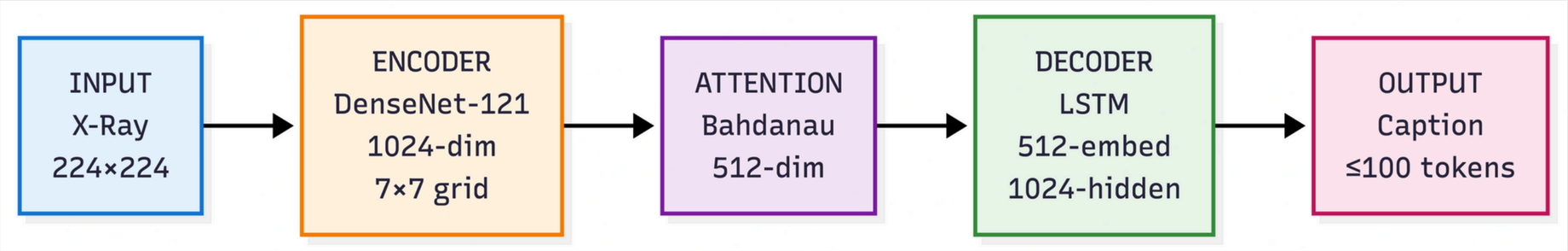


Example data split for first frontal strategy with min_freq=5

Model Architecture Overview

Model Statistics:

- Total parameters: ~22M
- Trainable parameters: ~15M
- Trainable percentage: ~68.0%



Training Setup & Hyperparameters

Hyperparameters:
Batch Size: 32 (auto-detectable)
Epochs: 30
Optimizer:

- type: “adam”
- learning rate: 0,0001

Early stopping:

- patience: 10
- metric: “val_loss”
- mode: “min”

Loss Function:

- Cross-entropy with no label smoothing
- Ignores <PAD> tokens in loss calculation

Data Augmentation:

- Rotation up to 5 degrees
- Color jitter:
  - brightness change up to 10%,
  - contrast change up to 10%,
- No horizontal flip (preserve anatomical laterality; impressions mention sides too)

Quantitative Evaluation

BLEU

(Bilingual Evaluation Understudy)

- Measures n-gram overlap with reference
- BLEU-1, BLEU-2, BLEU-3, BLEU-4
- Range: 0-1 (higher is better)
- BLEU-4 most commonly reported
- Good for: Word-level similarity

METEOR

(Metric for Evaluation of Translation with Explicit Ordering)

- Considers synonyms and stemming
- Aligns hypothesis and reference
- Better correlation with human judgment
- Range: 0-1 (higher is better)
- Good for: Semantic similarity

ROUGE-L

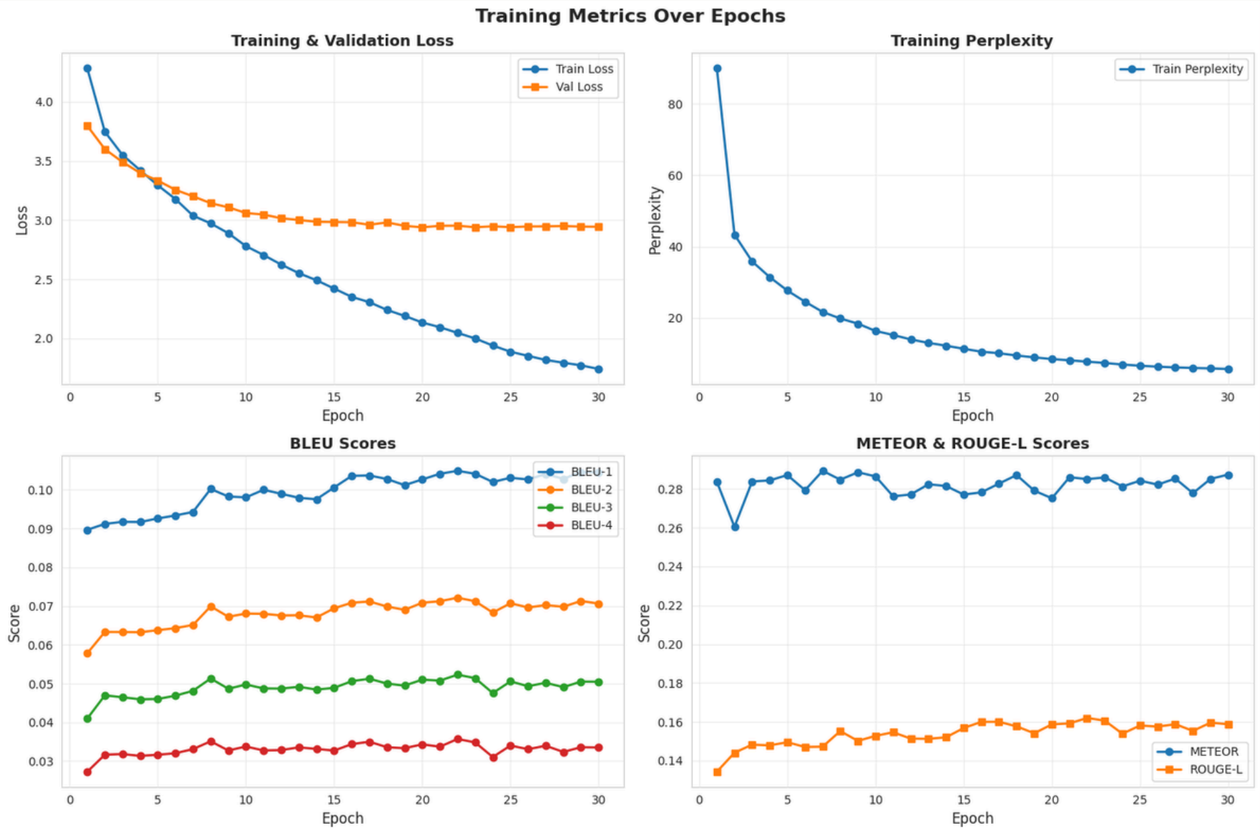
(Recall-Oriented Understudy for Gisting Evaluation)

- Longest common subsequence (LCS)
- Captures sentence-level structure
- Focuses on recall
- Range: 0-1 (higher is better)
- Good for: Sentence structure

Perplexity

- Measures model uncertainty
- Value of x means decision between x terms
- Lower is better
- Good for: Model confidence

Results - Metrics



Results using:

- first frontal strategy
- min_freq: 5

Epoch	Time (s)	Train Loss	Val Loss
20/30	113s	2.094	2.952

Perplexity	Bleu_1	Bleu2	Bleu_3	Bleu_4	Meteor	Rouge
8.203	0.104	0.071	0.051	0.034	0.286	0.159

Qualitative Analysis

Sample	Ground Truth	Generated
Sample A (UID: 123)	<UNK> right medial apical opacity which may be <UNK> to the <UNK> <UNK> upper lobe airspace disease or pulmonary nodule is not <UNK> <UNK> recommend <UNK> <UNK> chest and apical <UNK> view of the chest to further <UNK> findings and <UNK> were discussed <UNK> <UNK> in the <UNK>	no acute cardiopulmonary <UNK> <UNK> stable appearance of the <UNK> <UNK> deformity of the <UNK> <UNK> heart size is <UNK> <UNK> there is <UNK> no definite pleural effusion or <UNK> heart size is normal <UNK> no typical findings of pulmonary <UNK> heart size is normal and pulmonary <UNK> no <UNK>
Sample B (UID: 47)	vague opacity at the left lung base which appears to be within the left lower <UNK> this may represent <UNK> or <UNK> pneumonia given the <UNK> <UNK>	no acute cardiopulmonary <UNK> <UNK> stable appearance of the left lower lobe airspace disease with <UNK> <UNK> pleural <UNK> mediastinal contour within normal <UNK> no acute cardiopulmonary abnormality <UNK> stable mediastinal <UNK> mediastinal contour within normal <UNK> no acute cardiopulmonary abnormality <UNK> stable mediastinal <UNK> mediastinal contour within normal <UNK>

Empirical Error Analysis

1. **Vocabulary Limitations:**
   - Rare medical terms replaced with ‘<UNK>’
   - Solution: Increase corpus size or lower min_frequency
2. **Generic Descriptions:**
   - Produces safe but vague captions
   - “no acute findings” instead of specific details
3. **Multi-finding Cases:**
   - Difficulty describing multiple abnormalities
   - May focus on most prominent finding only
4. **Quantitative Assessments:**
   - Struggles with sizes (“mild vs moderate”)
   - Lacks precise location descriptions

Current Limitations

Dataset Limitations:

- Report quality varies across radiologists
- Vocabulary restricted to training corpus (total of 1497 tokens)

Methodology Limitations:

- Single-sentence caption (not full reports)
- No confidence scores for individual predictions
- Fixed input size (224x224) loses details
- Limited to chest X-rays

Next Steps & Improvement

Model Architecture Improvements:

- Experiment with Transformer decoder (vs current LSTM)
- Test different attention mechanisms
  - Currently implemented: Bahdanau, Luong
  - Future: Self-attention, Multi-head attention

Caption Quality Enhancements:

- Fix repetition loops in generated captions
  - Issue: Model repeats same phrases multiple times
  - Solution: Increase repetition penalty in beam search
- Address incorrect caption length generation
  - Issue: Captions too short or ignore length constraints
  - Solution: Length normalization and length reward

Multi-View Analysis:

- Compare captions for same patient across projections
  - Frontal vs Lateral caption consistency
  - Identify projection-specific vs common findings

THANK YOU

For Listening

Maria Linhares, 113534 Miguel Pinto, 107449