AutoImpress

Final Project Report

Team Members

Yaniv Grosberg

Netanel Ohev Shalom

Aviel Shmuel

Supervisor

Dr. Apartsin Alexander

AutoImpress – Clinical Impression Generation from Radiology Reports

📊 Data and Task

Dataset: IU-XRay (Indiana University Chest X-ray Reports)

- 3,955 reports with structured fields
- Fields: findings, indication, comparison, image, MeSH, Problems
- Free-text impression target

🎯 Task Definition

Input: Structured report fields

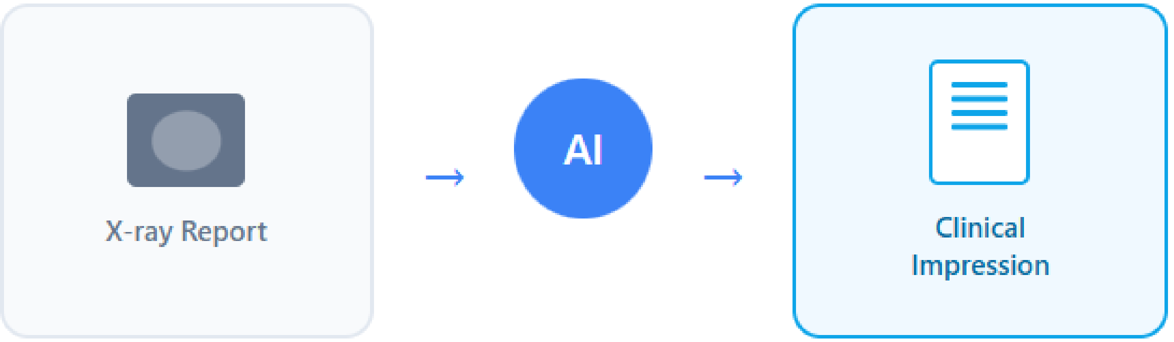
Output: Concise clinical summary

🤖 NLP Task Type

- Text-to-Text Generation
- Abstractive Summarization

📈 Evaluation Metrics

- LLM-based clinical equivalence scoring
- BERTScore (semantic similarity)



Prior Art

Source / Title	Approach / Model	Data	Metrics	Key Results
Zhang et al. (2023) *Leveraging Summary Guidance on Medical Report Summarization*	Fine-tunes BART/T5 on medical reports. Input: Findings + sampled example summary Output: Generated summary	MIMIC-III (DISCHARGE, ECHO, RADIOLOGY)	ROUGE, BERTScore, SummaC, QuestEval	Outperformed BART/T5 baselines, especially on DISCHARGE and ECHO
Ma et al. (2023) *From General to Specific: Domain Adaptation for Medical Report Generation*	Uses ChatGPT with prompts from similar cases, refined iteratively. Combines general + medical LLMs (HybridFusion). Input: Findings + similar reports Output: Impression section	MIMIC-CXR, OpenI	ROUGE, BERTScore	SOTA: FC-F1 = 80.09 (MIMIC-CXR); ROUGE-1 = 66.37 (OpenI); improved fluency and factuality.
Van Veen et al. (2023) *RadAdapt: Lightweight Domain Adaptation of LLMs*	Fine-tunes T5 with LoRA/Prefix for efficient impression generation. Input: Findings Output: Impression section (radiologist-validated)	MIMIC-CXR	ROUGE, BERTScore, human evaluation	Best performance with only 0.32% tuned params; clinically validated summaries.

Data Description & EDA

📋 Dataset: IU-XRay

Indiana University Chest X-ray Reports

📊 Dataset Statistics

- **Text length:** Findings ≈ 190, Impression ≈ 76 chars
- **Vocabulary:** ≈ 2,000 unique tokens
- **Templates:** "No acute cardiopulmonary abnormality" in 490+ records

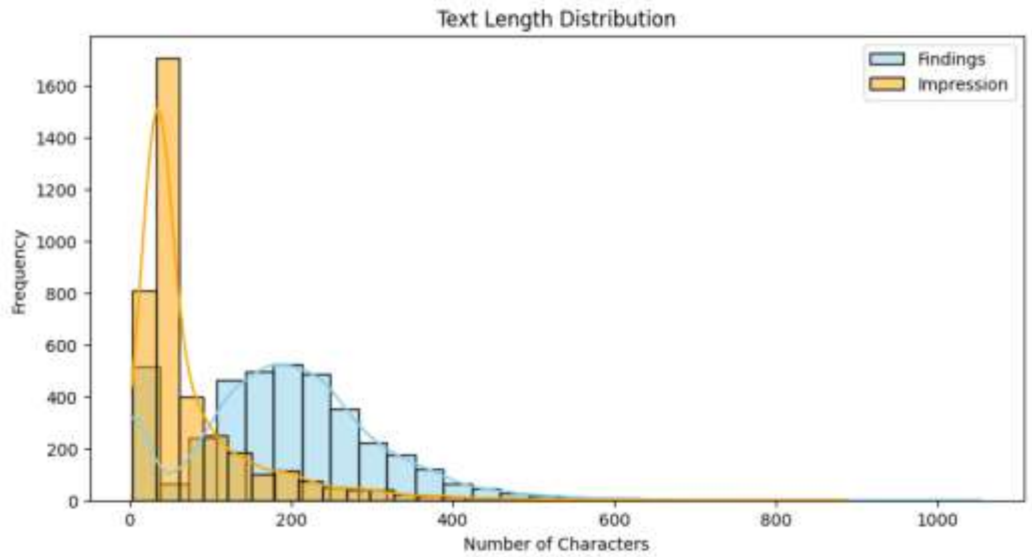
Phrase	Count
No acute cardiopulmonary abnormality	491
No acute cardiopulmonary findings	189
No acute cardiopulmonary abnormalities	168
No acute cardiopulmonary disease	163
No acute disease	126
No acute cardiopulmonary process	106
No acute radiographic cardiopulmonary process	93
No acute cardiopulmonary abnormality identified	80
No acute pulmonary disease	76
No acute findings	60

🔧 EDA Highlights

- Retained 3,331 records (removed ~520 incomplete)
- Filled missing fields with "none provided"
- Replaced anonymized patterns with [REDACTED]

💡 Key Insight

Short, highly templated clinical texts pose a challenge for true abstraction vs. surface-level copying.



Models & Processing Pipelines

🤖 Models Used

- **FLAN-T5** (google/flan-t5-base): Baseline + Fine-tuned
- **GPT-4.1** (Azure API): Inference & evaluation
- **DeepSeek-V3** (Azure API): Inference & evaluation

⚙️ FLAN-T5 Configuration

- **Epochs:** 3 | **Batch Size:** 4
- **Split:** 80% train / 20% test → 90% train / 10% validation
- **Platform:** Google Colab Pro (L4/A100 GPU)

🔄 Pipeline Overview

FLAN-T5:

Data → Few-shot → Baseline → Fine-tune → Validate → GPT-4o Judge

GPT-4.1 & DeepSeek:

Data → 50% Sample → API Generation → BERTScore → GPT-4o Judge

☁️ Inference Setup

- **Platform:** VSCode, Azure APIs
- **Computing:** Cloud-powered (no local GPU)

Metrics

🎯 Primary Metric

GPT-4o Judge: Binary clinical equivalence (YES/NO) comparing generated vs. ground truth impressions

📊 Secondary Metric

BERTScore (F1): Semantic similarity between generated and reference impressions

🔧 Computation Details

Training: Validation loss on 10% held-out set

Evaluation: BERTScore + GPT-4o judgment on test set

APIs: 50% sample for BERTScore + GPT-4o evaluation

💭 GPT-4o Judge Prompt Example

```
You are a medical expert. Compare the following two radiology
impressions. Determine if they are clinically equivalent in meaning,
and verify that the generated impression is written in appropriate
radiology language without including non-clinical prompt elements.
Reference Impression: {reference} Generated Impression: {generated}
If the generated text includes non-clinical formatting or prompt
tokens, consider it NOT clinically equivalent. Answer with "Yes" or
"No" only.
```

💡 Key Insight

GPT-4o judgments reflect clinical interpretability, while BERTScore provides surface-level semantic similarity.

Project Structure

GitHub Repository:

[View Project Repository]

Project Structure

```
project-root
│
├── data_raw/ → Raw datasets (indiana_reports.csv)
├── data_cleaned/ → Cleaned datasets (indiana_reports_cleaned.csv)
├── outputs/
│   ├── flan-t5-base/ → FLAN baseline & fine-tuned results
│   │   ├── generated_impressions_300_flan.csv
│   │   ├── finetuned_model_test_results.csv
│   │   ├── results_with_azure_gpt_judgment_baseline.csv
│   │   └── results_with_azure_gpt_judgment.csv
│   └── gpt-deepseek/ → GPT-4.1 & DeepSeek results
│       ├── gpt41_judged_results.csv
│       ├── deepseek_judged_results.csv
│       ├── gpt4_1_acute_findings_vs_ground_truth.csv
│       └── deepseek_acute_findings_vs_ground_truth.csv
│
├── notebooks/
│   ├── Preprocessing_EDA.ipynb → Exploratory analysis & cleaning
│   ├── flan-t5-base.ipynb → FLAN-T5 baseline & fine-tuning
│   ├── gpt_deepseek.ipynb → GPT-4.1 & DeepSeek evaluation
│   ├── Analyze_Results.ipynb → Final output analysis
│   └── utils_file.py → All shared utility functions
│
├── readme.md
└── requirements.txt → Python dependencies
```

File Types

- Folders
- CSV Files
- Notebooks
- Python
- Markdown
- Text

Outputs — Field Meaning

`uid`	unique identifier for the report
`generated_impression`	text generated by the model
`true_impression`	expert-labeled ground truth
`gpt_equivalence`	GPT-4o judgment ("Yes" or "No")

Intermediate/Baseline Results

📊 Baseline Model (FLAN-T5 Few-Shot)

- **GPT-4o Clinical Equivalence:** 1.3% (4/300 samples)
- **BERTScore F1:** 0.8382 ± 0.0281

→ Strong text similarity but very poor clinical equivalence

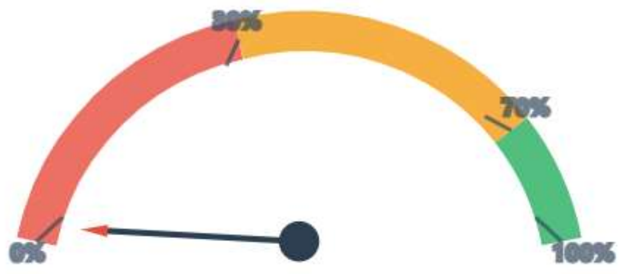
🔧 Training Metrics

- **Epochs:** 3, **Batch size:** 4
- **Final validation loss:** ≈ 0.1–0.2

⚙️ Intermediate Steps

- **Preprocessing:** removed ~520 incomplete records (~13% reduction)
- **Prompt engineering:** cleaned non-clinical tokens
- **Fine-tuning setup:** 80% train / 20% test, further 90% train / 10% validation

FLAN-T5 Clinical Equivalence Performance



1.3%

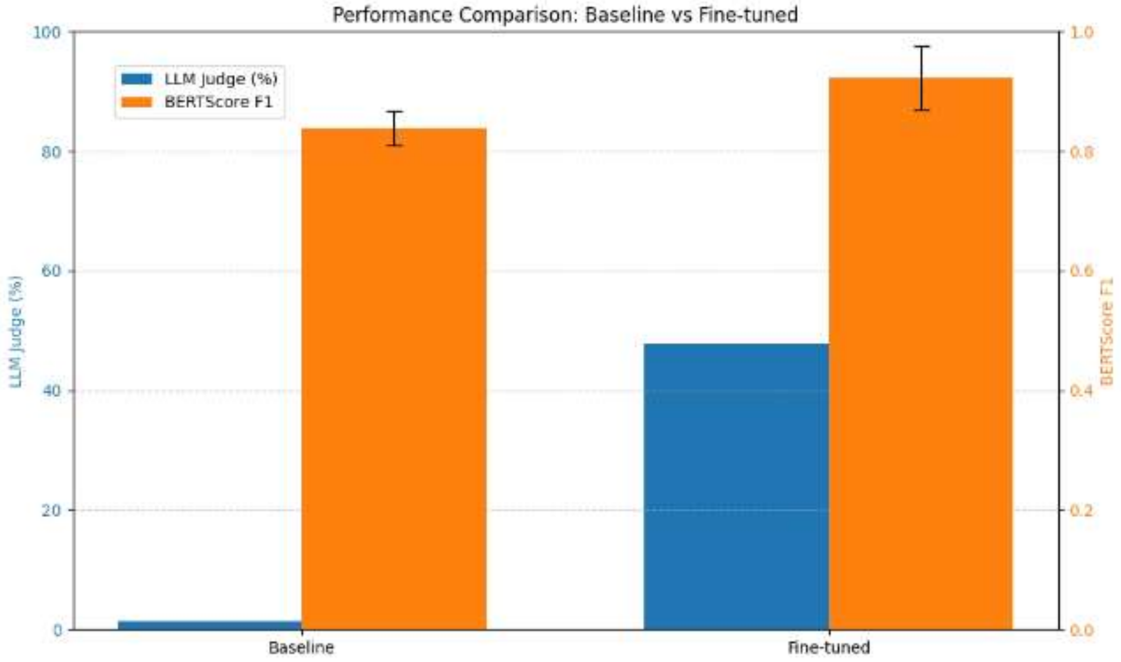
Clinical Equivalence Score

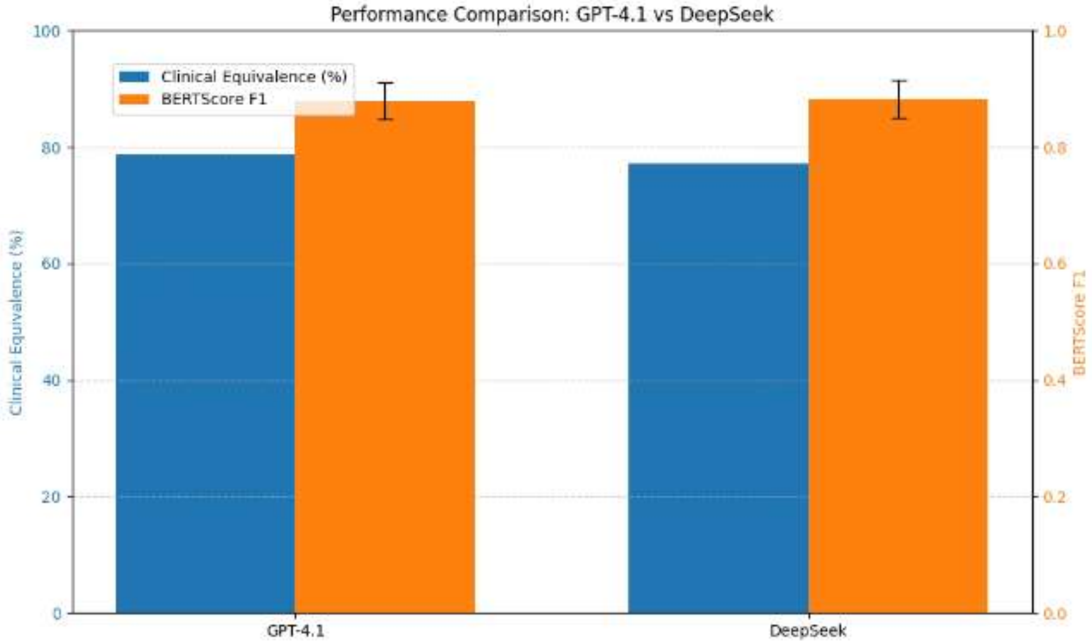
🔴 Poor Performance (0-30%) 🟠 Fair Performance (30-70%)

🟢 Good Performance (70%+)

Main Results

Model	GPT-4o Clinical Equivalence	BERTScore F1 (mean ± std)
FLAN-T5 Baseline	1.3% (4/300)	0.8382 ± 0.0281
FLAN-T5 Fine-Tuned	**47.7% (318/667)**	**0.9227 ± 0.0536**
GPT-4.1	**77.1% (1284/1666)**	0.8794 ± 0.0317
DeepSeek	**78.6% (1309/1666)**	0.8826 ± 0.0319





Conclusions

Fine-tuned FLAN-T5 showed substantial improvement over baseline, achieving major gains in both semantic similarity (BERTScore) and clinical equivalence (GPT-4o judgment).

GPT-4.1 and DeepSeek reached top-tier performance, providing a strong benchmark for large-scale, pretrained models without task-specific fine-tuning.

🎯 Prompt Design Impact

Multiple rounds of careful refinement were required to reach optimal formulation balancing clinical precision, language clarity, and minimal non-clinical artifacts.

📊 Dataset Patterns

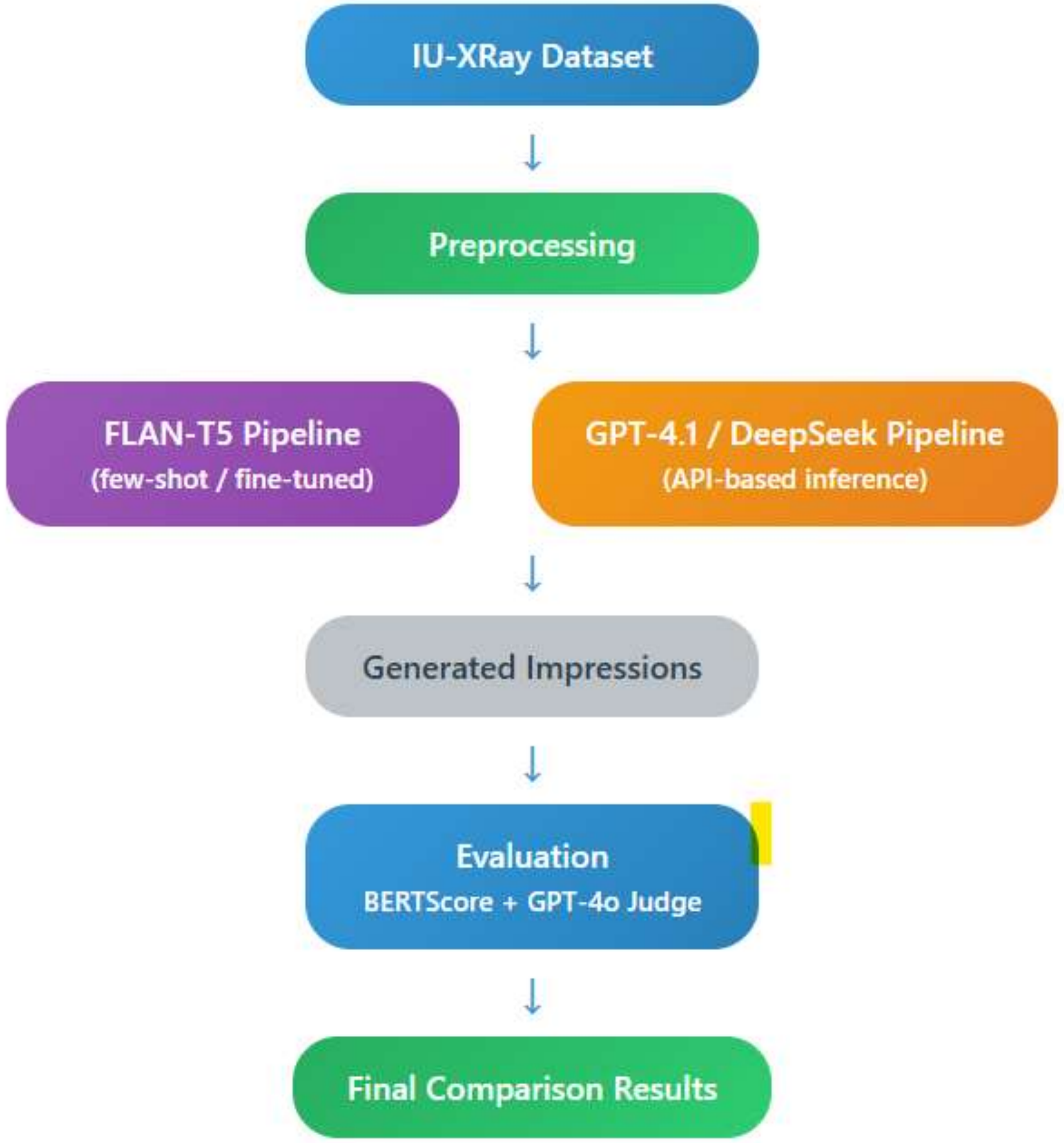
Advantage: easier to steer prompts toward realistic phrasing

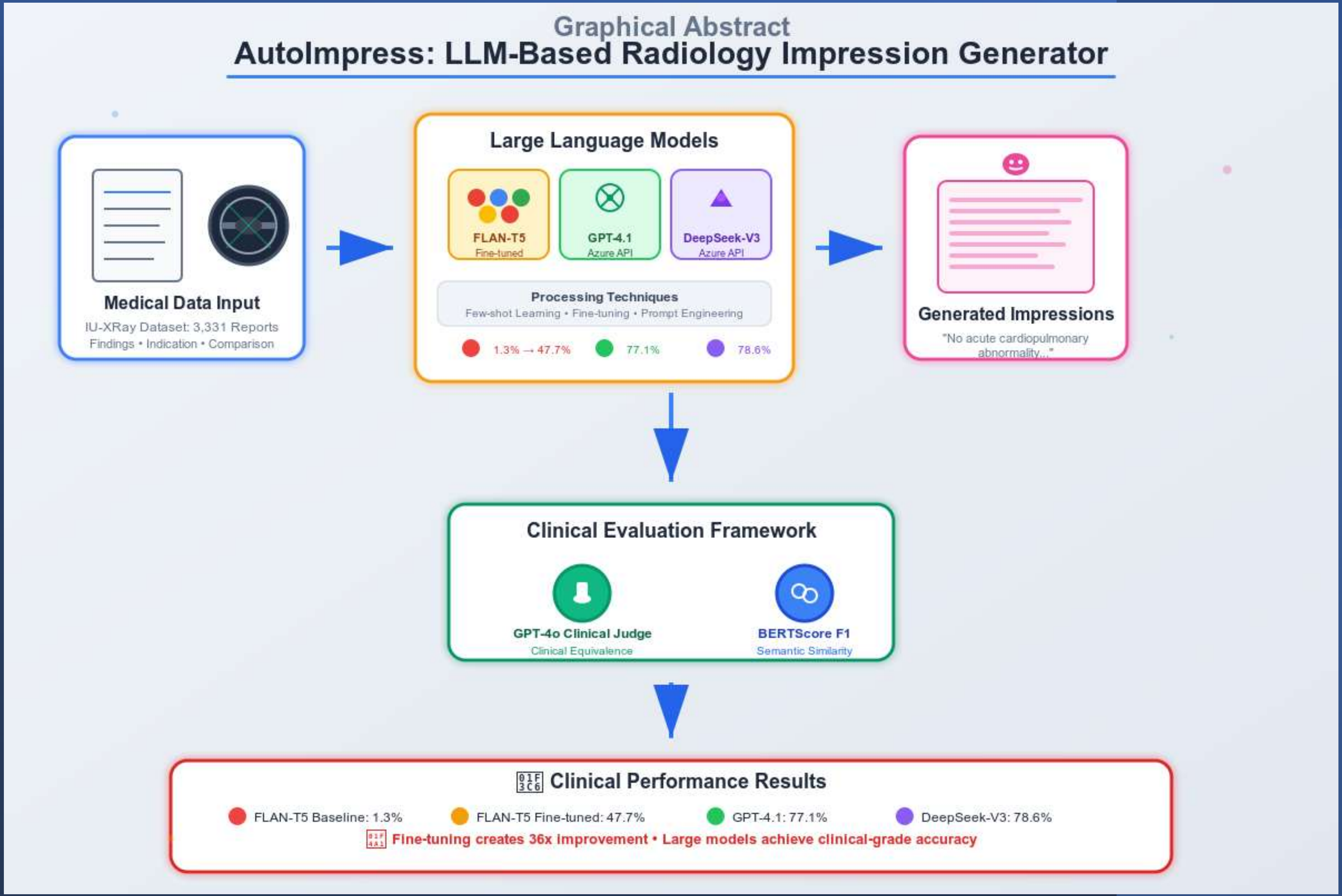
Constraint: limited generalization beyond common cases

🎉 Overall Impact

The project successfully met its objectives, demonstrating that structured input combined with fine-tuning and optimized prompting can substantially narrow the gap toward clinically meaningful text generation.

AUTOIMPRESS NLP Pipeline Architecture





Graphical Abstract
AutoImpress: LLM-Based Radiology Impression Generator
Medical Data Input
IU-XRay Dataset: 3,331 Reports
Findings • Indication • Comparison
Large Language Models
FLAN-T5
Fine-tuned
GPT-4.1
Azure API
DeepSeek-V3
Azure API
Processing Techniques
Few-shot Learning • Fine-tuning • Prompt Engineering
1.3% → 47.7%
77.1%
78.6%
Generated Impressions
"No acute cardiopulmonary abnormality..."
Clinical Evaluation Framework
GPT-4o Clinical Judge
Clinical Equivalence
BERTScore F1
Semantic Similarity
Clinical Performance Results
FLAN-T5 Baseline: 1.3%
FLAN-T5 Fine-tuned: 47.7%
GPT-4.1: 77.1%
DeepSeek-V3: 78.6%
Fine-tuning creates 36x improvement • Large models achieve clinical-grade accuracy