AutoImpress

Interim Project Report

Team members:

Yaniv Grosberg

Netanel Ohev Shalom

Aviel Shmuel

GitHub Repository: https://github.com/Yanivgg/AutoImpress

Project Overview

Project Title: AutoImpress – Clinical Impression Generation from Radiology Reports Using LLMs

Data and Task

- **Dataset:** IU-XRay (Indiana University Chest X-ray Reports) Includes 3,955 reports with structured fields (findings, indication, comparison, image, MeSH, Problems) and free-text impression.

Task

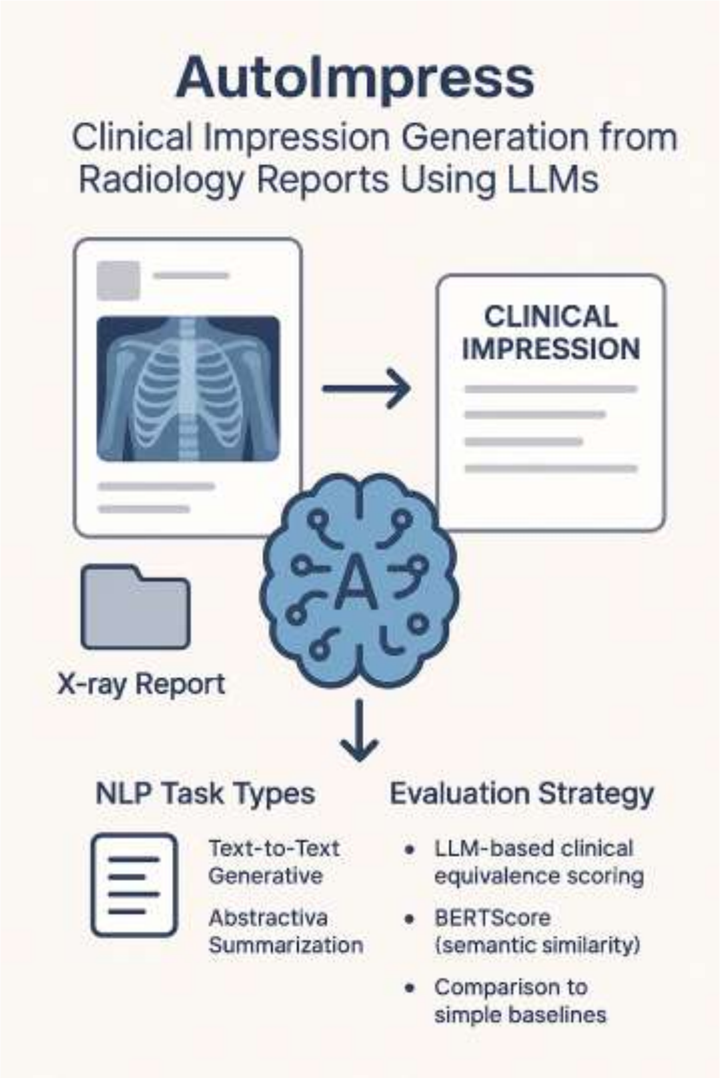
- **Input:** Structured report fields (findings, indication, comparison, image, MeSH, Problems)
- **Output:** Free-text impression – a concise, clinical summary written by the radiologist

NLP Task Type:

- Text-to-Text Generation
- Abstractive Summarization

Evaluation Strategy:

- LLM-based clinical equivalence scoring
- BERTScore (semantic similarity)
- Comparison to simple baselines



Slide 2 – Prior Art

Source / Title	Approach / Model	Data	Metrics	Key Results	Differences from Our Work
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	We use **structured field inputs** (Findings, MeSH, Indication) instead of full reports; and evaluate using LLM-based judgment, not just ROUGE
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.	Uses prompt retrieval and iterative refinement; we use direct prompting with a fixed template.
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.	Focuses on parameter-efficient tuning (LoRA, prefix); our work currently uses full model fine-tuning.

Pipeline & Plan

Stage	Input → Output	Approach	Evaluation
1. Preprocessing	Raw dataset → Cleaned text.	Token replacement, filtering missing entries	Row retention, text length stats
2. Prompt Design	Structured fields → Text prompt	Template-based few-shot prompting	Prompt completeness, token length.
3. Baseline Generation	Prompt → Generated impression	Pretrained LLM (e.g., FLAN-T5)	LLM Clinical Equivalence, BERTScore.
4. Baseline Analysis & Insights	Prompts + generations → Key observations	Analyzing the model outputs, identifying error patterns, and exploring potential optimization strategies"	Qualitative review of outputs, common phrase analysis, error categorization.
5. Fine-Tuning	Dataset + prior results → Fine-tuned or alternative model	Fine-tune T5 or medical LLMs (e.g., BioMedLM)	LLM Clinical Equivalence, BERTScore.
6. Model Improvement Analysis	Baseline vs. fine-tuned model	Performance delta analysis	Improvement in metrics

Exploratory Analysis & Baseline Evaluation

EDA & Preprocessing Summary

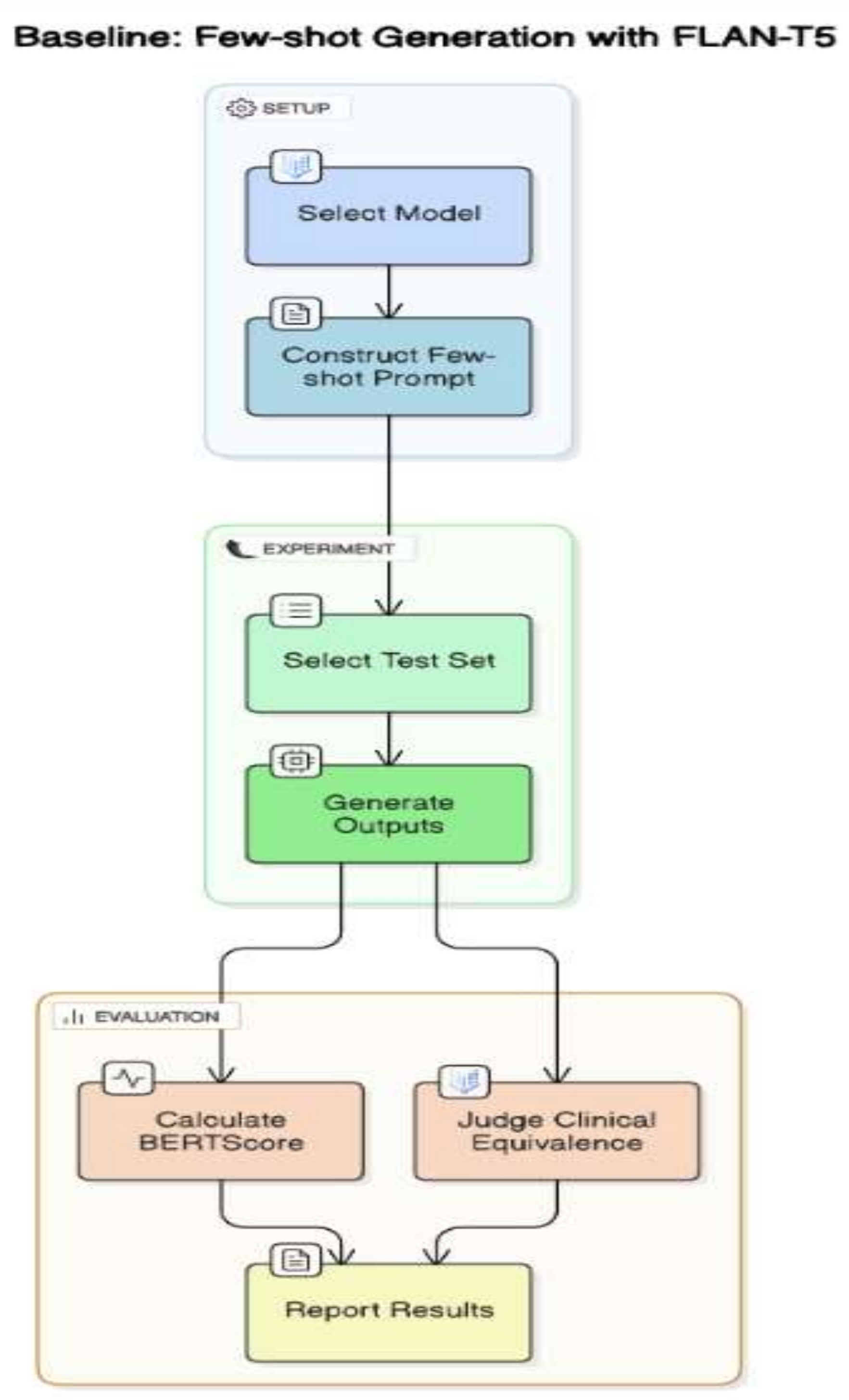
- Removed rows with missing impression or findings → **Rows before:** 3851, **Rows after:** 3331, **Removed:** 520
- Filled missing indication and comparison with "none provided"
- Replaced anonymized patterns (e.g., xxxx) with [REDACTED] to standardize inputs → Detected [REDACTED] in: **findings: 1425 rows**, **impression: 411 rows**
- Text length stats (characters): → findings: **mean = 190.6**, **std = 117.5** → impression: **mean = 76.2**, **std = 82.5**
- Frequent impression templates found (e.g., *"No acute cardiopulmonary…"*)

Baseline: Few-shot Generation with FLAN-T5

- **Model:** google/flan-t5-base
- **Approach:** Few-shot prompt using structured fields: image, indication, comparison, findings, MeSH, Problems
- **Test Set:** 100 samples

Evaluation Results:

- BERTScore (F1): 0.834 ± 0.024
- LLM Judge (GPT-4): Clinical Equivalence = 4 / 300 = 1.3%



Insights & Recommendations

Impressions are often short, repetitive, and follow templated phrasing (e.g., *"No acute cardiopulmonary findings"*)

About **40%** of impressions contain [REDACTED] tokens → impacts both training signal and evaluation clarity

BERTScore F1: 0.83, but **GPT clinical equivalence: only 1.3%** → indicates semantic gap between surface and clinical understanding

FLAN-T5 baseline shows copying tendencies from findings → lacks true abstraction or summarization

Next Steps

1 **Fine-tune a LLM model** to reach better performance (same as baseline or medical LLM) on this dataset to better capture domain-specific summarization patterns.

2 **Leverage impression templates**: cluster common phrasing patterns to inform guided generation.

3 **Evaluation Enhancements :**
1.Continue using **LLM-based clinical equivalence** as the primary evaluation method.
2.Use **BERTScore** for semantic similarity benchmarking, especially to assess improvements after fine-tuning.