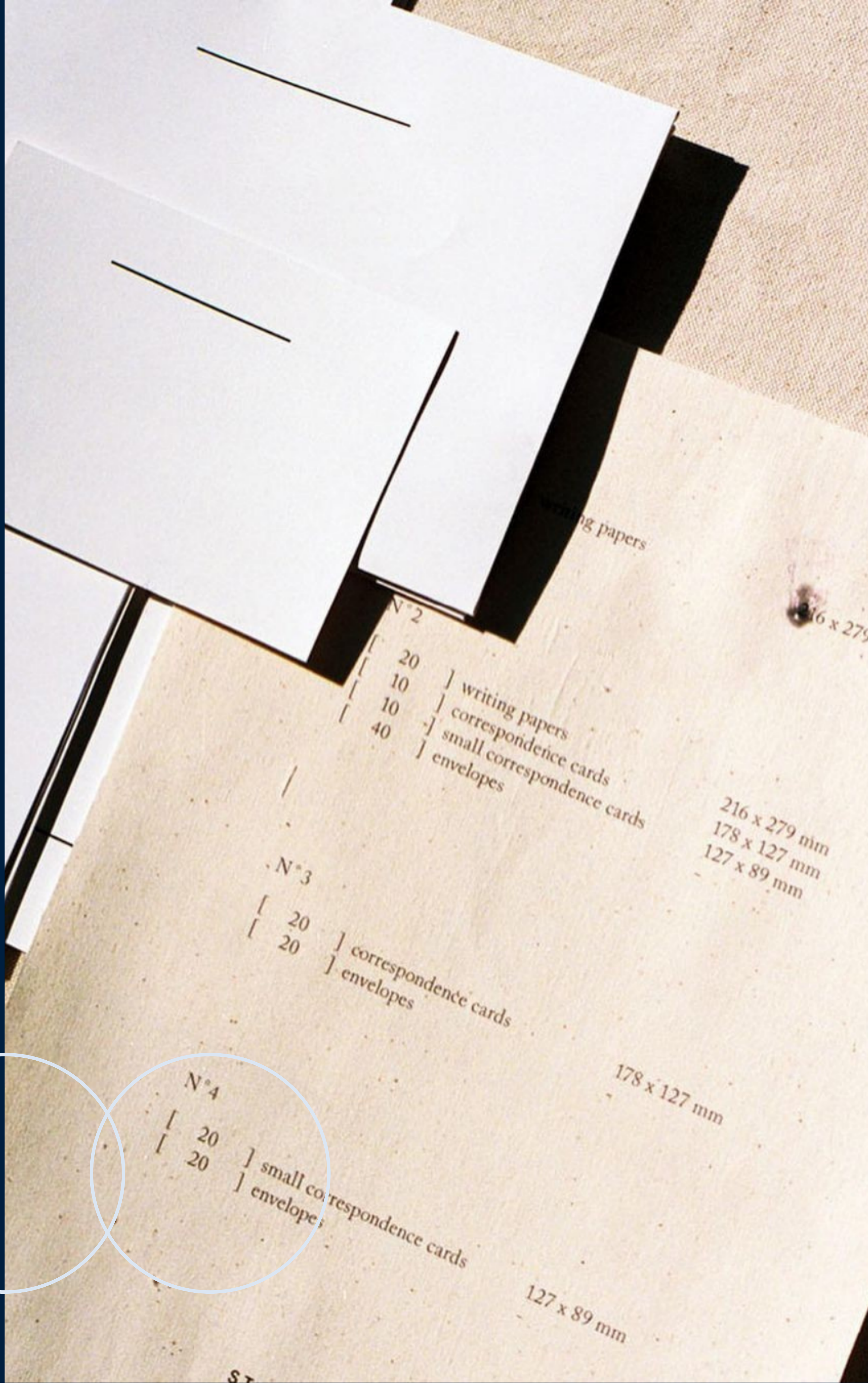
PROJECT 26

Summarization of Legal & Medical Documents

Gatri Reddy, Nishal Karamsetty, Deepak Reddy



Motivation

As a team, we observed that legal and medical documents are extremely long and difficult to read quickly. People working in these fields often struggle to extract key points under time pressure. We wanted to build a system that can simplify this process by generating short, clear summaries that save time and make information easier to understand.

Manual summarization is slow and inconsistent, especially when large volumes of documents are involved. An automated system that produces clear, accurate summaries can significantly improve decision-making and reduce cognitive load.

Existing summarization models struggle with legal and medical documents because of their length, complexity, and domain-specific language, leading to inaccurate or incomplete summaries.

Goal: build a unified summarizer that works for both legal bills and biomedical research.

Problem Statement

We aim to improve the slow and inconsistent process of reading and understanding long legal and medical documents. These texts are often dense, technical, and time-consuming for people to manually summarize.

Given a collection of legal bills or biomedical research articles, the goal is to generate a concise, clear, and meaningful summary that captures the essential information while keeping the original intent intact. Our focus is on producing reliable summaries across both domains using a single, efficient model.

Related Work

T5 Models

T5 models utilize a text-to-text framework, effectively transforming input documents across various tasks, showcasing versatility in language understanding.

BART & PEGASUS

BART combines bidirectional context with autoregressive decoding, while PEGASUS specializes in abstractive summarization with innovative pre-training strategies for enhanced performance.

Domain Models

Domain-specific models like LegalBERT, BioBART, and SciBERT optimize performance in legal and medical contexts, demonstrating advanced understanding of specialized terminologies.

Parameter-Efficient Tuning

Techniques such as LoRA, Prefix Tuning, and Adapters enhance model efficiency, enabling fine-tuning with fewer parameters while maintaining high performance across tasks.

Datasets Overview

BillSum (Legal)

BillSum is a dataset containing summaries of legal documents for training.

PubMed (Medical)

PubMed provides a comprehensive dataset of medical literature for analysis.

Combined Dataset

By training on balanced mixtures of both datasets, the model learns patterns from each domain, ensuring better generalization.

Methodology

Model Architecture

- Base model: T5-Small (can also use T5-Base if GPU allows)
- Fine-tuning strategy: LoRA (rank=8, α=32, dropout=0.1)
- Mixed-domain dataset: BillSum + PubMed combined

Pipeline

- Preprocess legal + medical datasets
- Tokenize text (512 input / 128 summary)
- Apply LoRA adapters to T5
- Train on mixed-domain dataset
- Evaluate using ROUGE metrics

Why LoRA?

- 90%+ fewer trainable parameters
- Works in Colab GPU
- Faster, cheaper, same (or better) accuracy

Experimental Design Overview

Model

LoRA-T5 (Fine-tuned on combined Legal and Medical datasets)

Constraint

Training limited to few epochs due to GPU limitations

Baseline Comparison

Lead-3 (Heuristic)
TextRank (Extractive)

Evaluation Metric

ROUGE Scores

Results Overview

LoRA-T5 Advantages

Produces meaningful summaries for long legal bills.
Accurately captures key ideas in medical texts.

Baseline Issues

Lead-3: Only picks initial sentences.
TextRank: Lacks context awareness.

Model	ROUGE-1	ROUGE-2	ROUGE-L
Lead-3	27.44	9.62	17.71
TextRank	31.51	10.08	18.46
T5 + LoRA (ours)	**28.04**	**10.86**	**18.86**

Qualitative Enhancements

LoRA-enhanced T5 models demonstrate significant improvements in summarization quality.

Analysis of LoRA

Impact and Future Directions

Efficient Fine-Tuning

By leveraging LoRA (Low-Rank Adaptation), we achieved significant model improvements without the constraints of powerful or expensive hardware, proving that resource-efficient training is viable.

Cross-Domain Generalization

The model successfully synthesized information from two highly distinct sectors—Legal and Medical. It bridged the gap between complex legalese and medical terminology within a single framework.

Competitive Advantage

This dual-domain capability sets our model apart. Most existing systems specialize in a single area, but our approach demonstrates that one system can effectively handle multiple, unrelated technical domains.

Team Contributions

Gatri Reddy

- Dataset preparation
- preprocessing
- evaluation.

Nishal Karamsetty

- Model training
- LoRA implementation
- debugging.

Deepak Reddy

- Baseline models
- ROUGE analysis
- report and presentation

Project Timeline

Phase 1

Initial research and selection of datasets for summarization tasks.

Phase 2

Development of the LoRA-enhanced T5 model architecture and training setup.

Phase 3

Conducting experiments to evaluate model performance and efficiency.

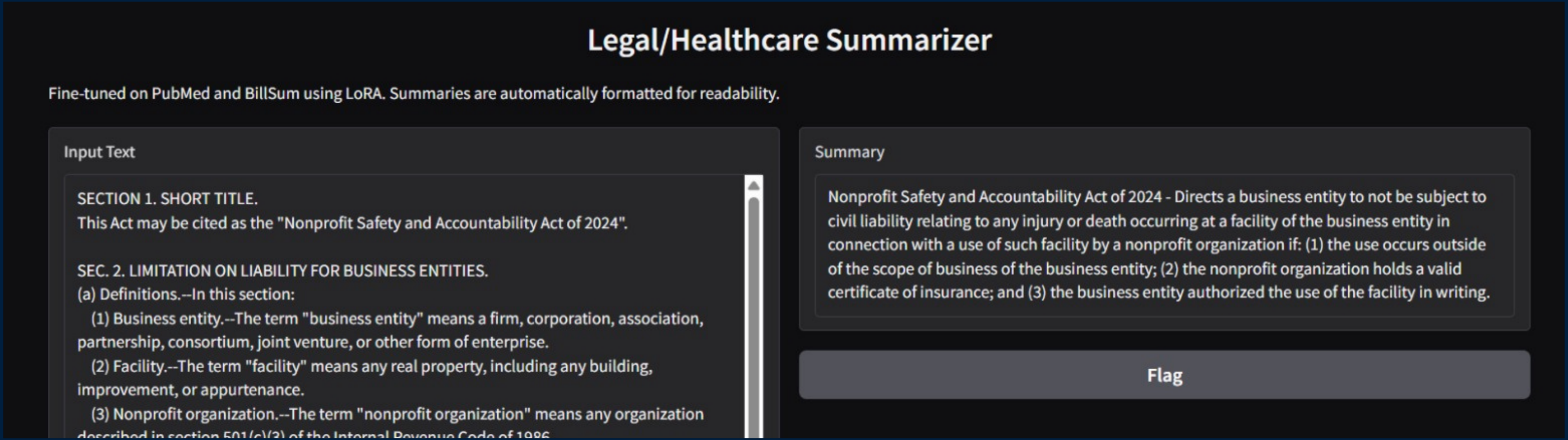
Phase 4

Analysis of results and comparison with existing summarization methods.

Phase 5

Finalizing project documentation and preparing for presentation of findings.

Demo & Application



Sample Inputs

Inputs are formatted for efficient processing by the summarization model.

Generated Summaries

Summaries effectively condense information while retaining essential details.

Baseline Comparison

Comparisons demonstrate improvements over traditional summarization methods.