Artificial Intelligence

NEWSGLANCE

News Summarisation

Group no. 11

Our Team

Radhika Bhati - 220650
Deepanshu Aggarwal - 220660
Rohit - 220657

TABLE OF CONTENT

PROBLEM STATEMENT

01

Overwhelming Amount of News Articles

02

Time-Consuming to Read Full articles

03

Busy Lifestyles (Lack of Time)

04

Need for Automated Summarization

05

Clear, Concise, and Accurate Summaries

OBJECTIVES

- **Automate Summarization:** Develop a system to summarize news articles into concise texts of 60-100 words.
- **Use Dual Techniques:** Implement both extractive (key sentences) and abstractive (new sentences) summarization methods.
- **Utilize Large Dataset:** Base the system on the CNN/DailyMail dataset with over 300,000 articles and summaries.
- **Save Time:** Provide quick news access and contribute to NLP and deep learning advancements.

BRIEF OVERVIEW OF APPROACHED APPLIED

We have applied two Extractive and two Abstractive Approaches

01 Frequency-Based Method

- Preprocessing
- Tokenization
- Word Frequency Calculation
- Sentence Scoring and Ranking
- Summary Generation

02 TF-IDF Based Method

- Data Preprocessing
- TF-IDF Vectorization
- Sentence Scoring
- Summary Generation

03 RNN-Based Method

- Encoder
- Attention Mechanism
- Decoder
- Training Pipeline

04 T5 Based Method

- Tokenizer
- Encoder
- Decoder
- Fine-Tuning

EXTRACTIVE APPROACHES

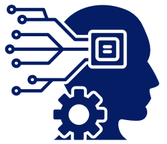
Frequency Based Method

- Preprocessing: Remove stop words, punctuations, and convert text to lowercase.
- Tokenization: Split the text into sentences and words.
- Word Frequency Calculation: Compute normalized frequency of each word in the text.
- Sentence Scoring and Ranking: Score sentences based on the sum of normalized word frequencies and rank them.
- Summary Generation: Select top-ranked sentences to form the summary.

TF-IDF Based Approach

- Data Preprocessing: Clean text by removing noise and tokenize into sentences.
- TF-IDF Vectorization: Compute term frequency (TF) and inverse document frequency (IDF) for word weights.
- Sentence Scoring: Score sentences by summing TF-IDF weights of words.
- Summary Generation: Select sentences with the highest scores for the summary.

ABSTRACTIVE APPROACHES



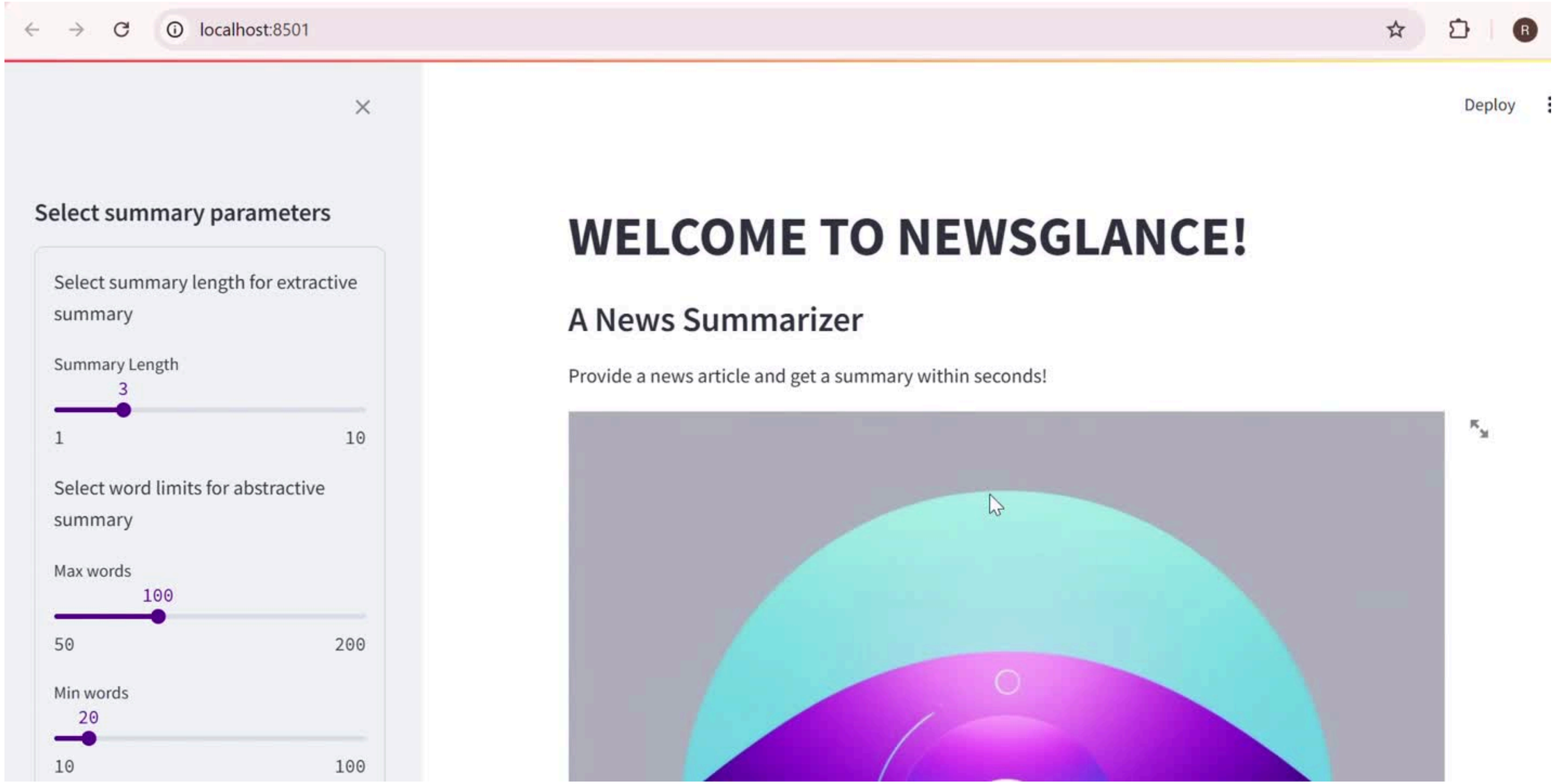
RNN Based Method

- Encoder: Convert tokens to dense representations using pre-trained GloVe embeddings. Use a bidirectional GRU to encode the input sequence.
- Attention Mechanism: Focus on relevant parts of the input during decoding.
- Decoder: Generate summary tokens using GRU and project output through softmax.
- Training Pipeline: Employ teacher forcing and gradient clipping for optimized learning.

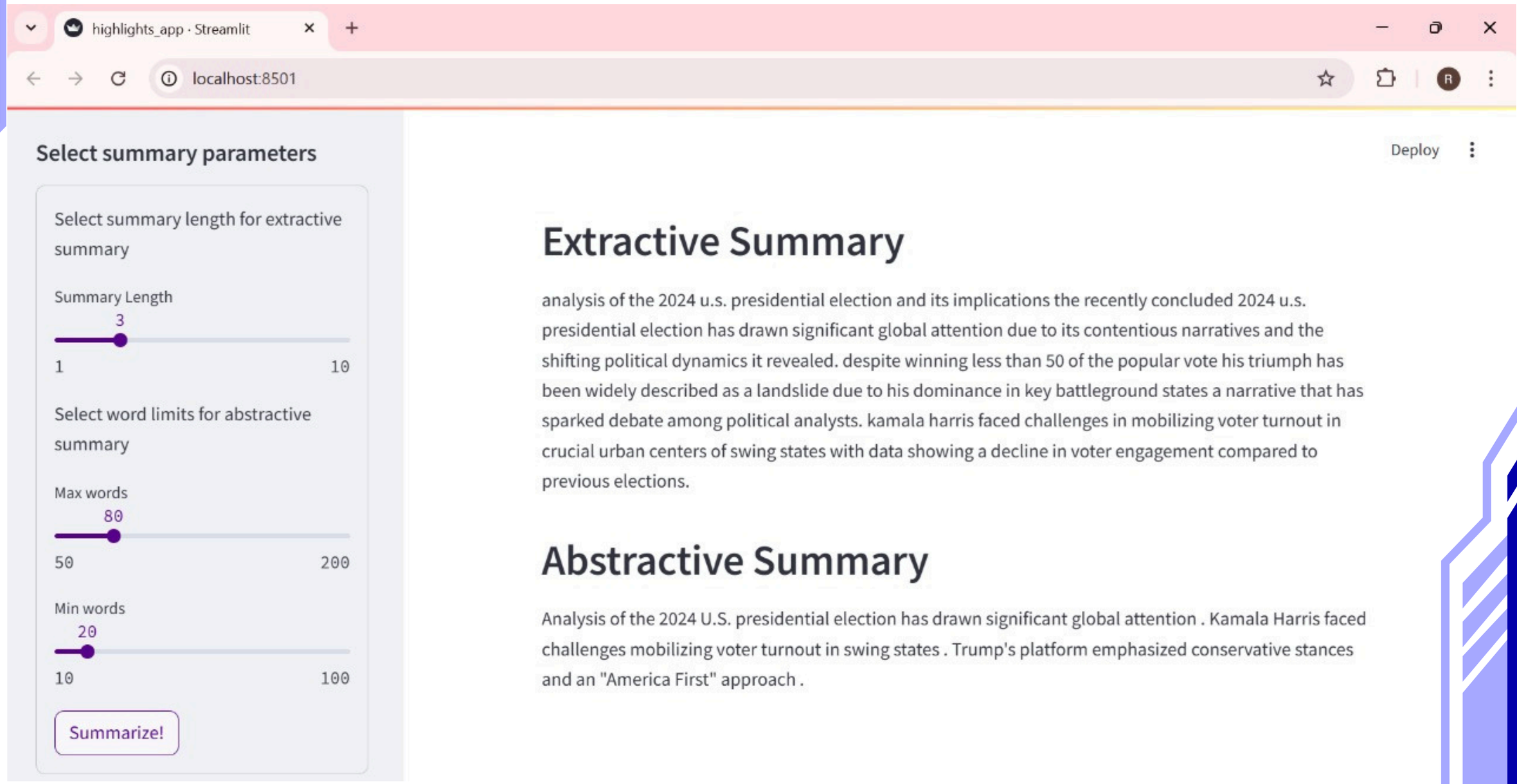
T5 Based Approach

- Tokenizer: Convert text into subword tokens and manage padding/truncation for batching.
- Encoder: Use multi-head self-attention and feed-forward layers to encode context.
- Decoder: Generate summaries sequentially using cross-attention on encoder outputs.
- Fine-Tuning: Adapt pre-trained T5 for summarization using task-specific data.

DEMONSTRATION



DEMONSTRATION



RESULTS

• For extractive summarization, the TF-IDF-based approach is the best, providing high accuracy and relevance in extracted sentences.

• T5 models outperform RNNs significantly for abstractive summarization, providing coherent, semantically rich summaries.

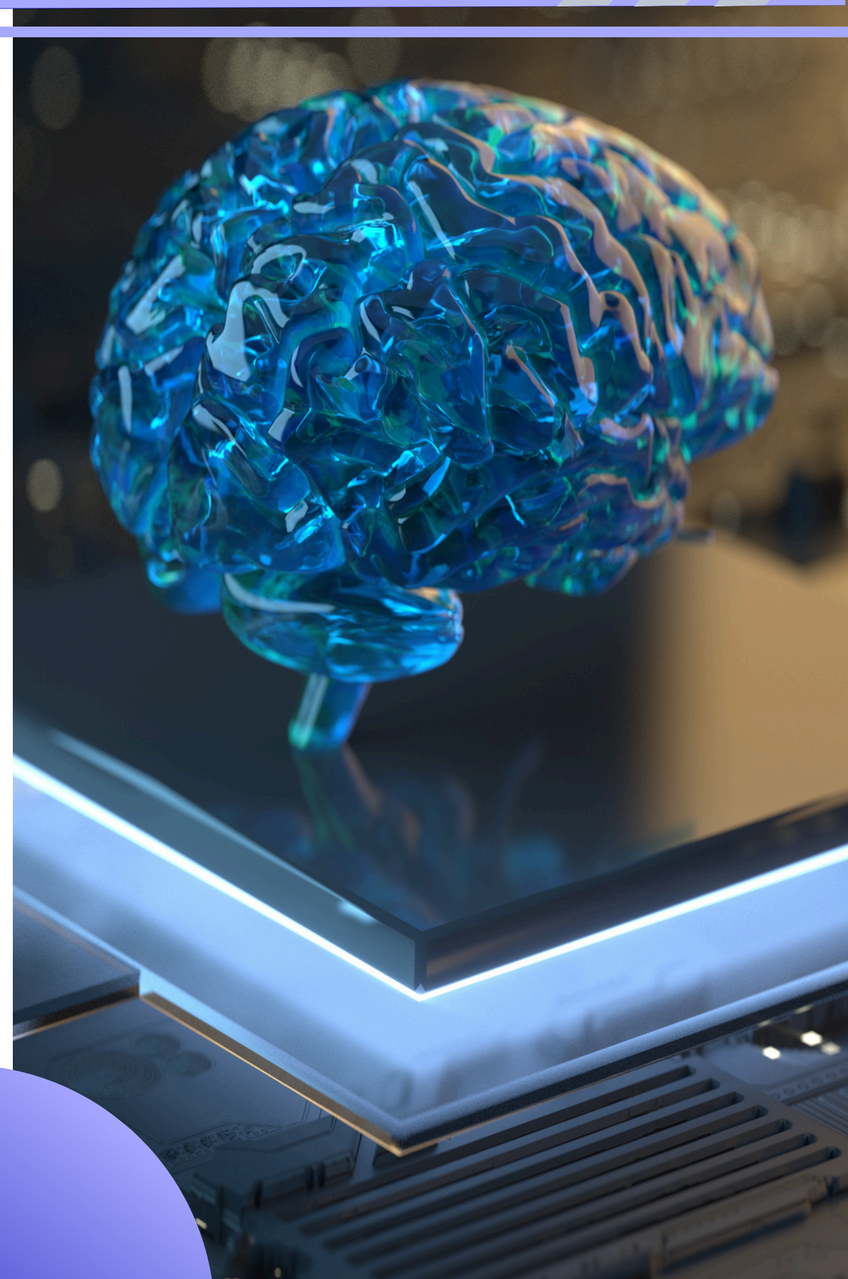
Metric	Text ranking based	TF-IDF based	Best Approach
ROUGE-1	15.97	35.08	TF-IDF
ROUGE-2	8.3	16.7	TF-IDF
ROUGE-L	11.87	23.16	TF-IDF
METEOR	33.38	37.53	TF-IDF
BERTScore	81.09	86.69	TF-IDF

Metric	RNN	T5	Best Approach
ROUGE-1	14.32	27.07	T5
ROUGE-2	2.21	9.16	T5
ROUGE-L	12.01	19.69	T5
METEOR	9.23	21.31	T5
BERTScore	75.06	86.38	T5

CONCLUSION

Both extractive and abstractive summarization methods are effective but serve different purposes. Extractive methods excel in preserving original information, while abstractive methods, like the T5 transformer, create more human-like summaries.

Importantly, these methods are not directly comparable. The results underscore that no single approach is universally superior; the choice depends on the use case and requirements, such as precision, creativity, or contextual relevance.



FUTURE SCOPE

- **Hybrid Models:** Combine extractive and abstractive methods for better accuracy and readability.
- **Domain-Specific Fine-Tuning:** Tailor models to specific fields for improved performance.
- **Enhanced Metrics:** Develop new evaluation metrics for better semantic and coherence assessment.
- **Multilingual Summarization:** Expand to multiple languages for global applicability.

THANK YOU