

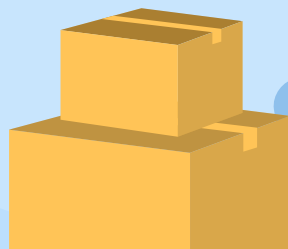
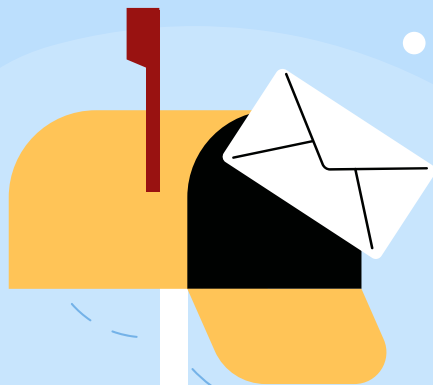
# Email Summarization

LGTM: Abhishek, Mrinoy, Samyukta



# Table of contents

- 01. **Data Pipeline**
- 02. **Data**
- 03. **Models**
- 04. **Evaluation**
- 05. **Demo**
- 06. **Future Growth**



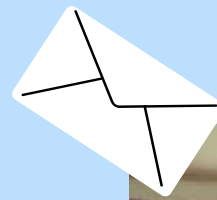
# Motivation: Reclaim your time

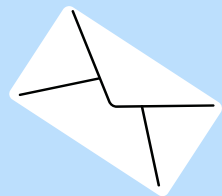
The Email Quagmire:

Average individuals spend **2+ hours** daily sifting through emails. Missed emails burgeon into daunting threads, consuming even more precious time.

Major Concerns:

1. Time it takes to catch-up
2. Missing mails



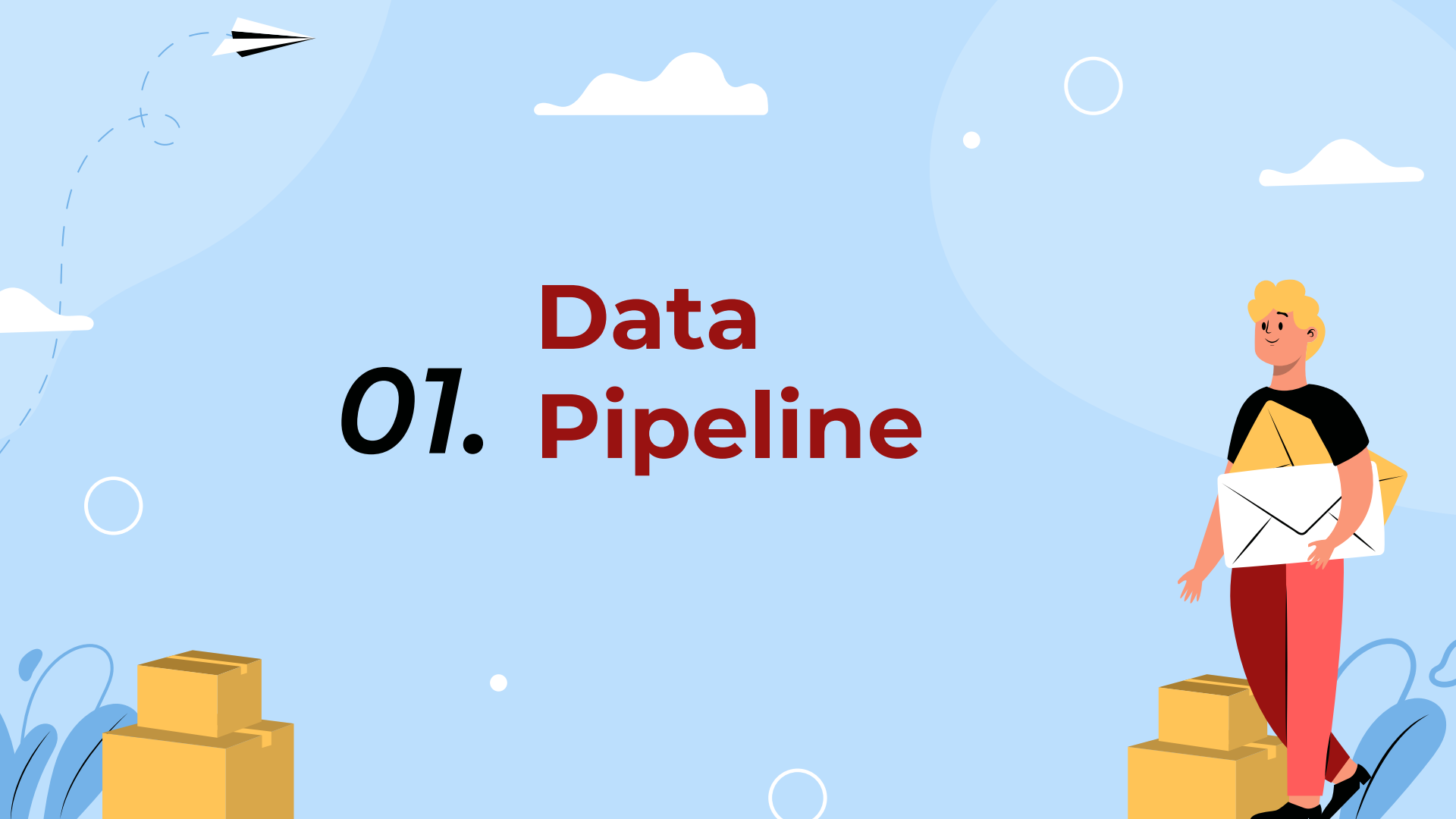


# Literature Review

Paper: EMAILSUM: Abstractive Email Thread Summarization (2021)

| EmailSum Long   | rouge1 | rouge2 | rougeL | BERTScore |
|-----------------|--------|--------|--------|-----------|
| T5 base         | 43.41  | 13.81  | 29.97  | 31.58     |
| HT5             | 43.86  | 14.06  | 30.17  | 31.84     |
| Semi-sup. (w3c) | 43.99  | 14.18  | 30.56  | 32.04     |

# 01. Data Pipeline





**UI: Chrome  
Extension**

**Credentials**



**Fast API  
Server**

**Email summaries**



**Emails stripped  
of HTML**

**Summaries**



**Fine-tuned  
BART model**

**3 layer unfreeze  
fine tuning**

**kaggle**

**Dataset (4157  
Email Threads)**

## 02. Data



# Merged Dataset

kaggle

From: Theriot, Kim S.

Sent: Tuesday, January 29, 2002 1:23 PM

To: Richardson, Stacey; Anderson, Diane; Gossett, Jeffrey C.;

....

Cc: Aucoin, Evelyn; Baxter, Bryce; Wynne, Rita

Subject: FW: Master Termination Log

Attached is the Daily Termination List for January 25 as well as the Master Termination Log, which incorporates all terminations received through January 25.

The following were previously on the Master Termination Log and have now been marked as "Y" for a valid

.....

From: Panus, Stephanie

Sent: Thursday, January 31, 2002 12:08 PM

To: Adams, Laurel; Albrecht, Kristin; ..

Subject: Master Termination Log

Attached are the Daily Lists for January 29 and January 30 as well as the Master Termination Log, which incorporates all terminations received through January 30. Also, prepetition mutual terminations have been added to this list. ..


## Summary:

The email thread discusses the Master Termination Log and the need to investigate a CNG LDC (Hope Gas) termination and a \$66 million settlement offer. Stephanie Panus sends out the Daily List and Master Termination Log for various dates. Kim Theriot requests her name and Melissa Murphy's name to be removed from the distribution list and adds several names to it. The thread also includes updates on terminations and valid terminations for various companies.

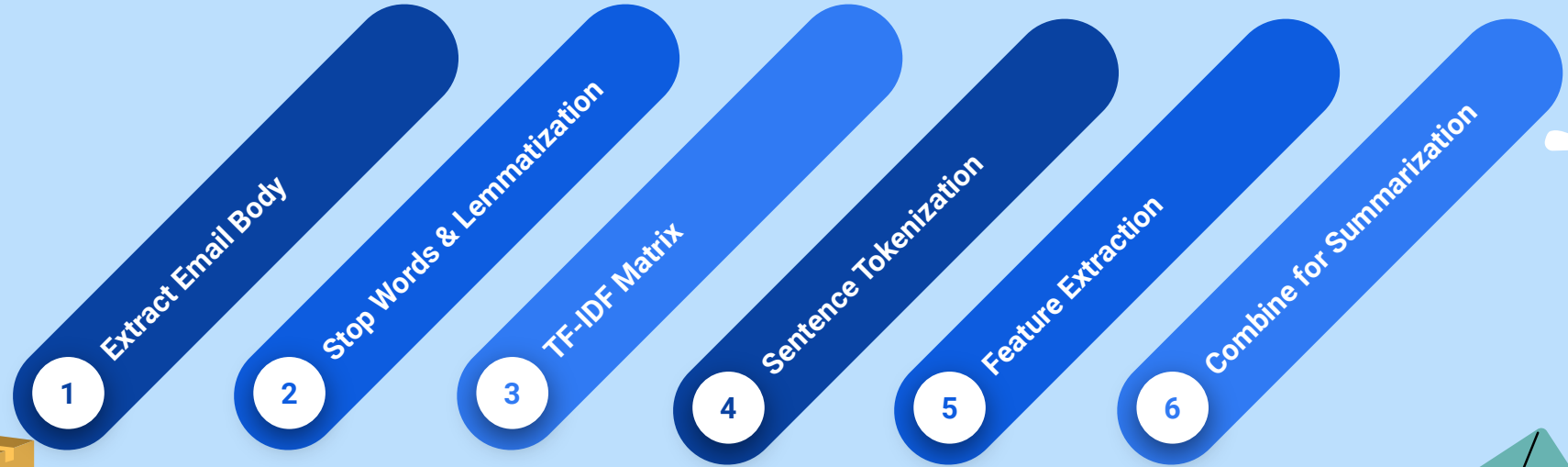


## 03. Models

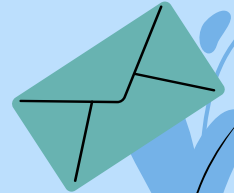




# Non-Deep Learning Approach: TF-IDF



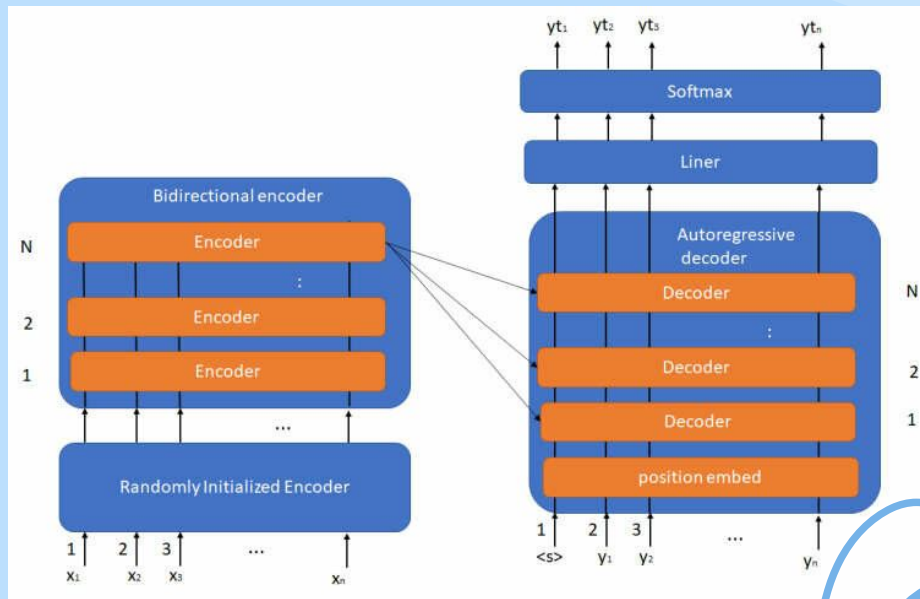
**Issues:** Coherence - Extractive Summarization



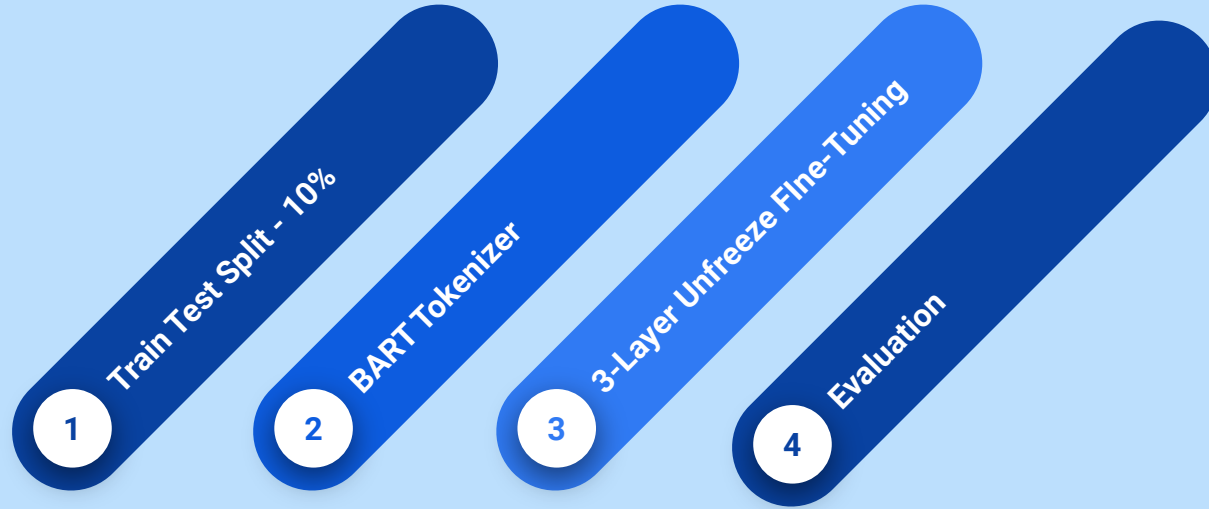


# Bidirectional and Auto-Regressive Transformers

1. Utilizes a bidirectional encoder (like BERT) to deeply understand text context.
2. Employs an autoregressive decoder (like GPT) to generate coherent and contextually relevant text.
3. Simple Model : 1024 context window & 139M parameters



# Bidirectional and Auto-Regressive Transformers



**Advantage:** Coherence - Abstractive Summarization

# 04. Evaluation





reference

$\ell_{ref}^{unigram} = 10$

→ The way to make people trustworthy is to trust them

hypothesis

$\ell_{hyp}^{unigram} = 9$

→ To make people trustworthy, you need to trust them

## Rouge-1

- Measures the overlap of unigrams (single words).
- Presence of individual words

## Rouge-2

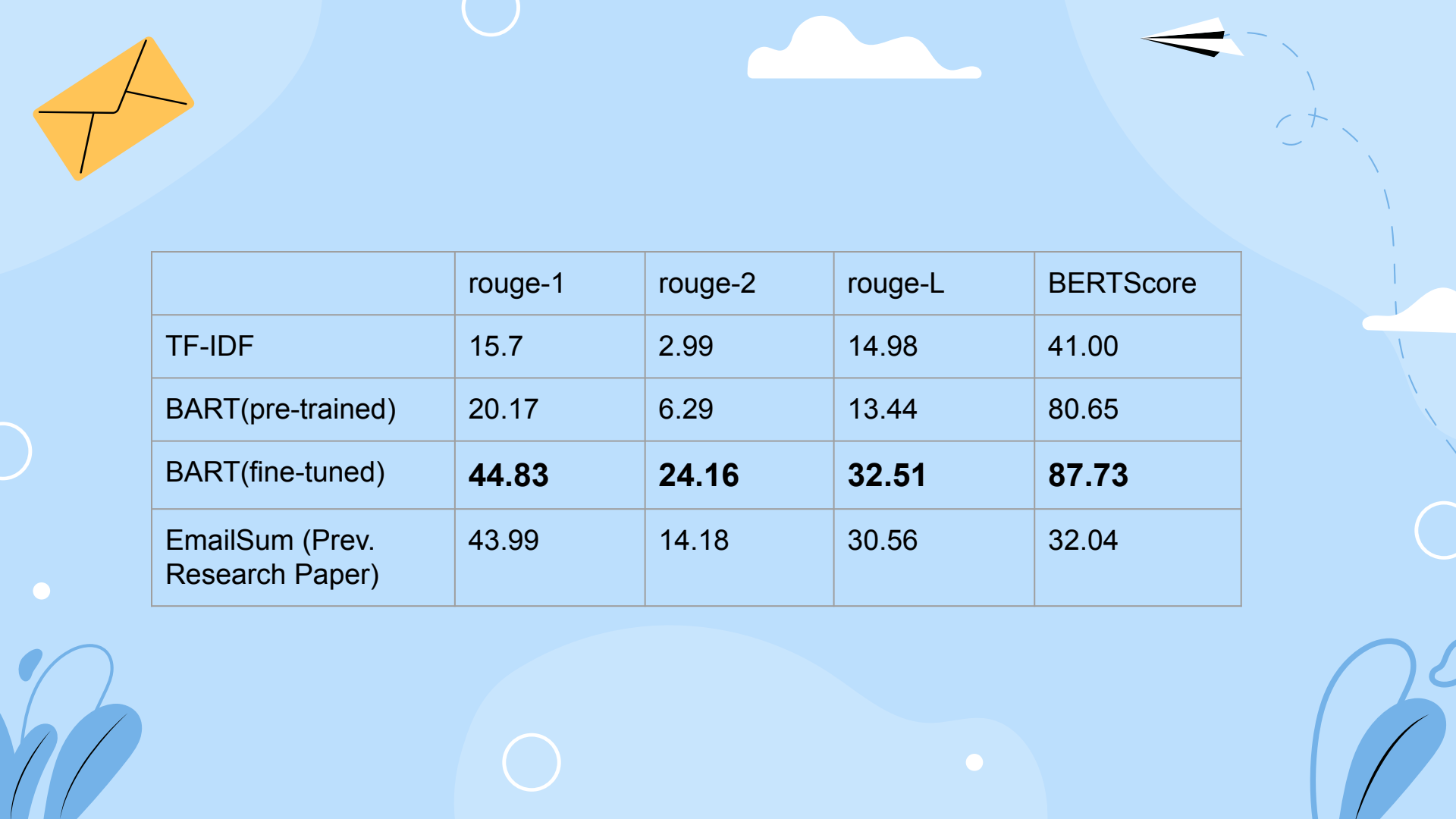
- Measures the overlap of bigrams (pairs of consecutive words).
- Preservation of word pairs

## Rouge-L

- Assesses the longest common subsequence.
- Evaluating the overall structure and order of the content

## BERTScore

- Considers semantic similarity, making it more robust to paraphrasing and variations in word choice.



|                                 | rouge-1      | rouge-2      | rouge-L      | BERTScore    |
|---------------------------------|--------------|--------------|--------------|--------------|
| TF-IDF                          | 15.7         | 2.99         | 14.98        | 41.00        |
| BART(pre-trained)               | 20.17        | 6.29         | 13.44        | 80.65        |
| BART(fine-tuned)                | <b>44.83</b> | <b>24.16</b> | <b>32.51</b> | <b>87.73</b> |
| EmailSum (Prev. Research Paper) | 43.99        | 14.18        | 30.56        | 32.04        |

# 05. Demo







## 06. Future Growth

1. We plan to use a more robust model with a larger context window (llama2, Mistral).
2. Encrypt and store credentials and/or use OAuth.
3. Application basis:
  - a. Summarise mails multiple accounts.
  - b. Configure summary email's frequency.

# LGTM! Thank You!

Any  
questions?

