ECE1786 - Project Final Report

EmailSense: Email Classifier and Summarizer using GPT4-Turbo

Isabella Hao 1004960633

Weizhou Wang 1004421262

Word count: 1794 + 52 (Table 1 content) + 14 (footnote) + 138 (Appendix A Email content) - 6(intext reference) = 1992

(excluding cover page, titles, figures, tables, labels, and references)

Penalty: 0%

Permissions

Team Member: Weizhou Wang

Permission to post video: yes
Permission to post final report: yes
Permission to post source code: yes

Team Member: Yicong (Isabella) Hao

Permission to post video: yes
Permission to post final report: yes
Permission to post source code: yes

1.0 Introduction

By 2023, email overload had become a significant issue, with workers receiving an average of 65.5 emails daily [1]. University professors face even greater challenges, often sorting through hundreds of academic emails[2][3].

After consulting with multiple university professors, we developed EmailSense[1]. This GPT4-Turbo-powered email Classifier and Summarizer is designed to reduce email burdens for professors by categorizing emails into five types: curriculum-related (CURR), project collaborations (COLLAB), reference letter requests (RECOM), administrative communications[2] (ADMIN), and others (OTHER). EmailSense further generates summaries for emails that warrant attention, helping professors prioritize responses and enabling swift overviews. The system also integrates with Apple Focus for better accessibility and user control.

2.0 Illustrations

As illustrated in Figure 1, EmailSense is triggered when a user activates a preset Focus mode on their device. This system includes a monitoring engine that tracks user's incoming emails via the IMAP server. Each incoming email is then individually analyzed by a GPT4-Turbo-based Classifier and Summarizer. The Classifier categorizes emails into one of five categories, and the Summarizer creates one-sentence summaries for time-sensitive emails.

When Focus mode is deactivated, another Shortcut activates the Result Generator, which compiles the categorized and summarized data into a comprehensive activity report, subsequently sent back to the user's inbox. An output example is provided in Appendix A.

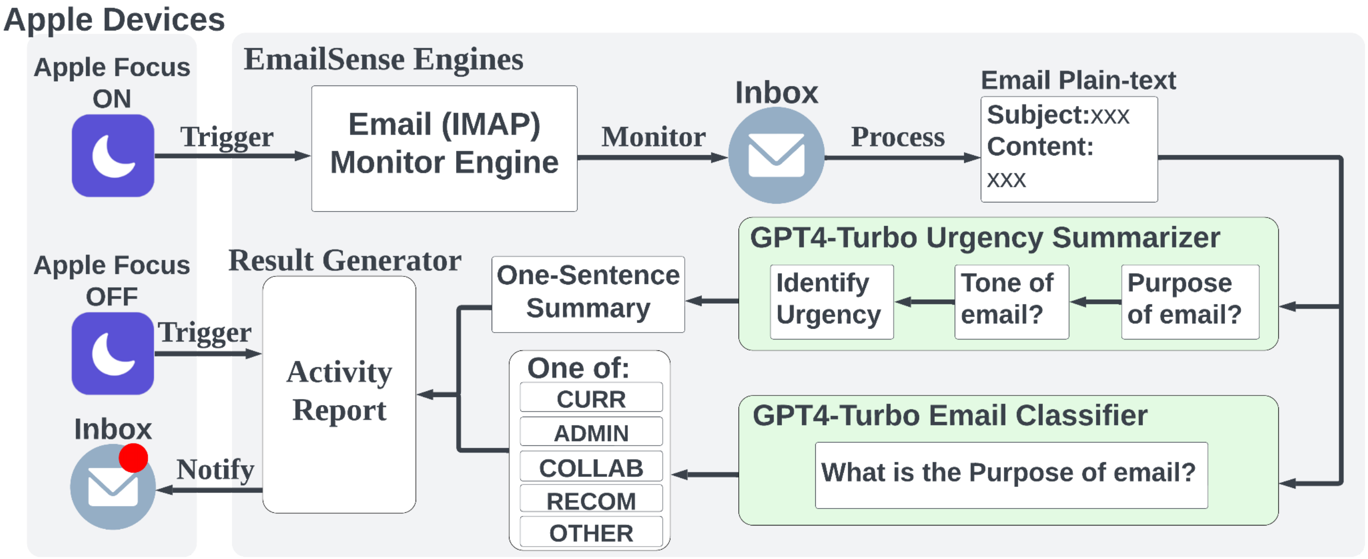


Figure 1. System Block Diagram

[1] Source code, prompts and demo: https://github.com/ece1786-2023/EmailSense

[2] Administrative tasks such as recruitment, budget management, etc.

3.0 Background & Related Works

Email classification and summarization have garnered substantial attention in Machine Learning over the years.

In 2023, a study found that Large Language Models (LLMs), including GPT-4, effectively detect phishing in human-written emails [4]. The results showed that the accuracy increased when the LLMs were specifically prompted to identify suspicious content, in comparison to general intention, highlighting the importance of precise prompting in LLMs.

In 2021, another paper developed an Email thread summarization dataset consisting of 2,549 email threads [5]. The dataset was tested on various models, including BertSumExt and T5, and the results showed that T5 and its variants outperformed the rest. The research also noted a disconnect between automated metrics (ROUGE, BERTScore) and human evaluations, stressing the significance of human judgment in assessing summarization quality.

4.0 Data & Data Processing

Our team developed a semi-synthetic dataset to train and test the email Classifier and Summarizer. For each of the five categories, 10 real emails from online resources and our academic mailboxes were initially gathered. Next, using the GPT4-Turbo model and tailored prompts, we generated 40 synthetic emails for each category, culminating in a balanced dataset of 250 emails. We manually ensured all emails were realistic and varied in length, tone, and subject, and targeted university professors.

In assessing the Summarizer, our goal was to ensure that 10% of the emails in our dataset were time-sensitive. Initially, only eight emails met this criterion. To increase this number, as illustrated in Table 1, we employed another prompt to inject time-sensitive information into additional emails. This approach augmented our dataset with 20 more time-sensitive emails, culminating in a total of 28.

Finally, the dataset was randomly split into 80% of the validation set (200 emails) and 20% of the hold-out set (50 emails) with the distribution shown in Tables 2 and 3.

Before Injection	**After Injection**
Could you clarify this for us during the next lecture?	As the homework is due tomorrow, could you clarify this for us as soon as possible?
Your input will be invaluable, particularly regarding the new laboratory course proposals.	It is crucial that all faculty members should provide immediate feedback regarding the proposal.

Table 1. Before-Injection vs. After-Injection for Selected Email Excerpts (Word Count Included)

	CURR	**COLLAB**	**RECOM**	**ADMIN**	**OTHER**
Real Emails	8	7	6	8	8
Synthetic	31	37	32	32	31
TOTAL	**39**	**44**	**38**	**40**	**39**

Table 2. Statistics of real emails and synthetic emails in 200 training samples.

	CURR	**COLLAB**	**RECOM**	**ADMIN**	**OTHER**
Real Emails	2	3	4	2	2
Synthetic	9	3	8	8	9
TOTAL	**11**	**6**	**12**	**10**	**11**

Table 3. Statistics of real emails and synthetic emails in 50 hold-out samples.

5.0 Architecture & Software

As illustrated in Figure 2, our models extracted each email's subject and content into plain text, which is analyzed by the Classifier and Summarizer. Both models have been carefully prompted following an instruction-format-example structure[6].

The Classifier interprets the email's intention and aligns it with predefined categories. For straightforward categories, one-sentence definitions are sufficient, while for broader categories like ADMIN, we provided a list of possible topics to guide the model, enhancing its comprehension. In the output, beyond just categorizing, the Classifier articulates its understanding of the email's intention, thereby solidifying its predictions.

The Summarizer employs a chain-of-thought approach to mimic manual urgency detection. In the prompt, we first defined urgent emails as requiring the professor's attention by the next day. Then, we asked the model to determine the email's purpose, assess the tone and word choices, and eventually identify the time sensitivity. In the end, its output includes an urgency label, a summary for time-sensitive emails, and a chain-of-thought explanation.

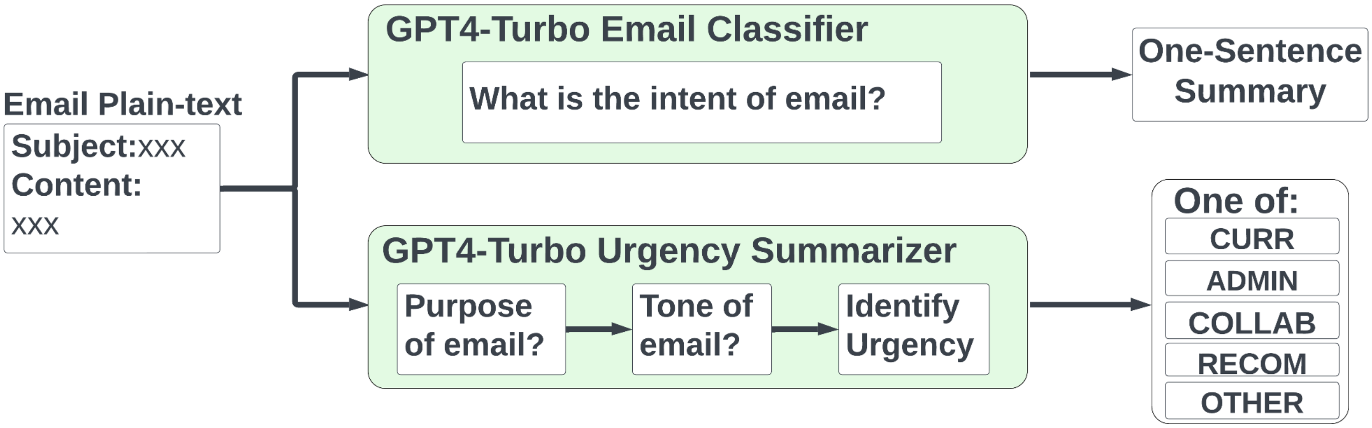


Figure 2. Neural Network Model Architecture

6.0 Comparison

To evaluate the Classifier, we employed a confusion matrix to assess potential biases and performance in each category. Since each category is equally important, a Macro-F1 Score was used for evaluating the model's overall performance.

For the Summarizer evaluation, we focused on accurately summarizing time-sensitive emails and avoiding including non-urgent emails in summaries. This process can be streamlined as a binary classification task, using a confusion matrix and F1 Score for performance assessment. Additionally, a manual review ensured the summarized content aligned with the original email's intent and critical information.

The same evaluation was also performed on GPT3.5-Turbo models with the same prompts as baseline comparisons.

7.0 Quantitative Results

This section presents the quantitative analysis of the Classifier and Summarizer's results and the comparison with respective baseline models.

7.1 Classifier Quantitative Results

The classifier achieved an overall accuracy of 99%, correctly classifying 198 out of 200 validation samples. As depicted in Figure 3-left and Table 4, there was only one real email and one synthetic email from COLLAB and ADMIN that were mislabeled as OTHER. As shown in Table 5, these outcomes result in full F1 scores for RECOMM and CURR emails, while at least 0.975 scores for the other three categories. The overall Macro F1 score is 0.990.

In comparison, as illustrated in Figure 3 right, the baseline generated worse results in almost all categories. Particularly in OTHER, about 40% of emails were mistakenly classified as ADMIN. As shown in Table 6, this underperformance resulted in a low Recall of 0.487 and an overall

Macro-F1 score of 0.861. Notably, as depicted in Table 4, almost 70% of its misclassifications were from synthetic emails.

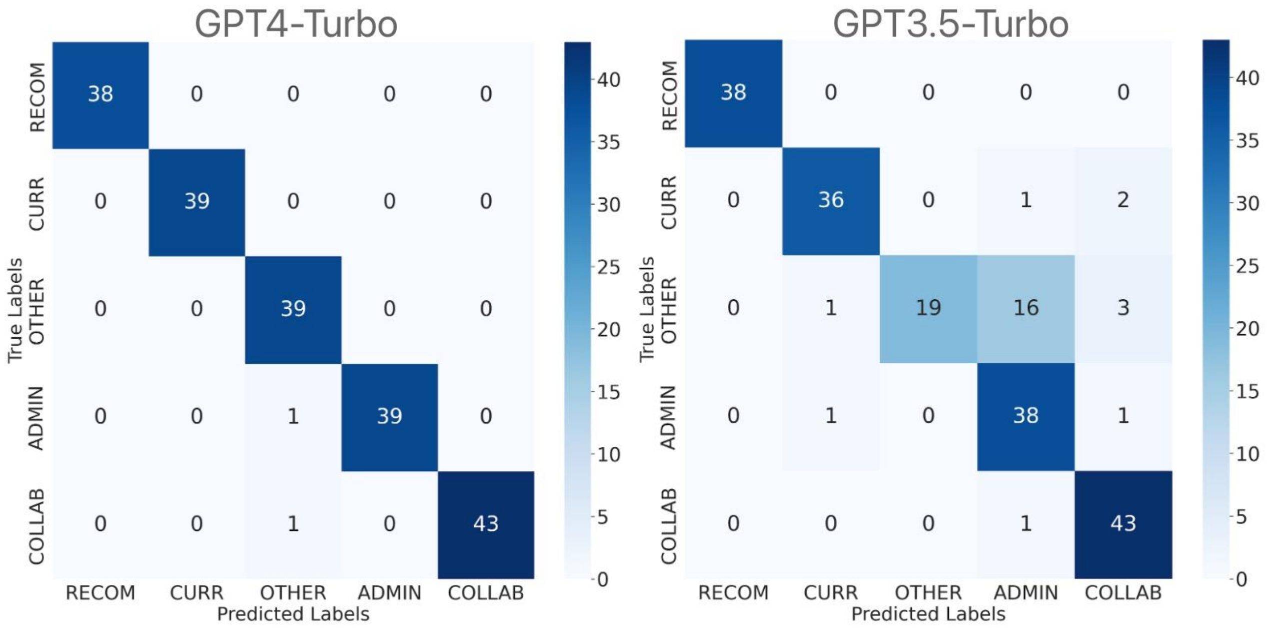


Figure 3. Validation Confusion Matrices of GPT4-Turbo models (left) vs. GPT3.5-Turbo baseline models (right) on Classification

		Misclassified Real Email	**Misclassified Synthetic Email**
Classifier	**GPT3.5-Turbo Validation**	8	18
	GPT4-Turbo Validation	1	1

Table 4: Real vs. Synthetic Data Performance on Classifiers

Category	**Precision**	**Recall**	**F1 Score**
RECOM	1.000	1.000	1.000
CURR	1.000	1.000	1.000
OTHER	0.951	1.000	0.975
ADMIN	1.000	0.975	0.987
COLLAB	1.000	0.977	0.989
Macro Average	**0.990**	**0.990**	**0.990**

Table 5: Validation Performance Metrics for GPT4-Turbo Models on Classification

Category	Precision	Recall	F1 Score
RECOM	1.000	1.000	1.000
CURR	0.947	0.923	0.935
OTHER	1.000	0.487	0.655
ADMIN	0.679	0.950	0.792
COLLAB	0.878	0.977	0.925
Macro Average	**0.901**	**0.867**	**0.861**

Table 6: Validation Performance Metrics for GPT3.5-Turbo Baseline Models on Classification

7.2 Summarizer Quantitative Results

In terms of urgency identification, the baseline model (Figure 4-right) demonstrated high accuracy in identifying non-urgent items with only four misclassifications. However, among 23 urgent emails, the baseline neglected 7 time-sensitive ones, corresponding to a 0.696 Recall and an overall F1 score of 0.744 from Table 7. Moreover, similar to the Classifier results, almost all misjudgements occurred in synthetic emails (see Table 8).

In contrast, the validation model (Figure 4-left) showed a remarkable improvement, especially in the identification of urgent emails with perfect Recall, indicating no urgent items were missed. Meanwhile, there were only three misclassifications (two real emails and one synthetic email) for non-urgent emails, suggesting fewer false positives. These improvements gave the model a 0.939 F1 score, significantly higher than the baseline.

In evaluating the summarization quality, our manual review revealed that all 26 summaries created by GPT4-Turbo effectively captured the intended messages of the emails and accurately identified key details such as times and locations. In contrast, the baseline model, while accurately capturing the emails' main objectives, produced considerably shorter summaries. This brevity resulted in 35% (7 out of 20) of summaries omitting crucial deadlines, potentially impacting users' perception of urgent information.

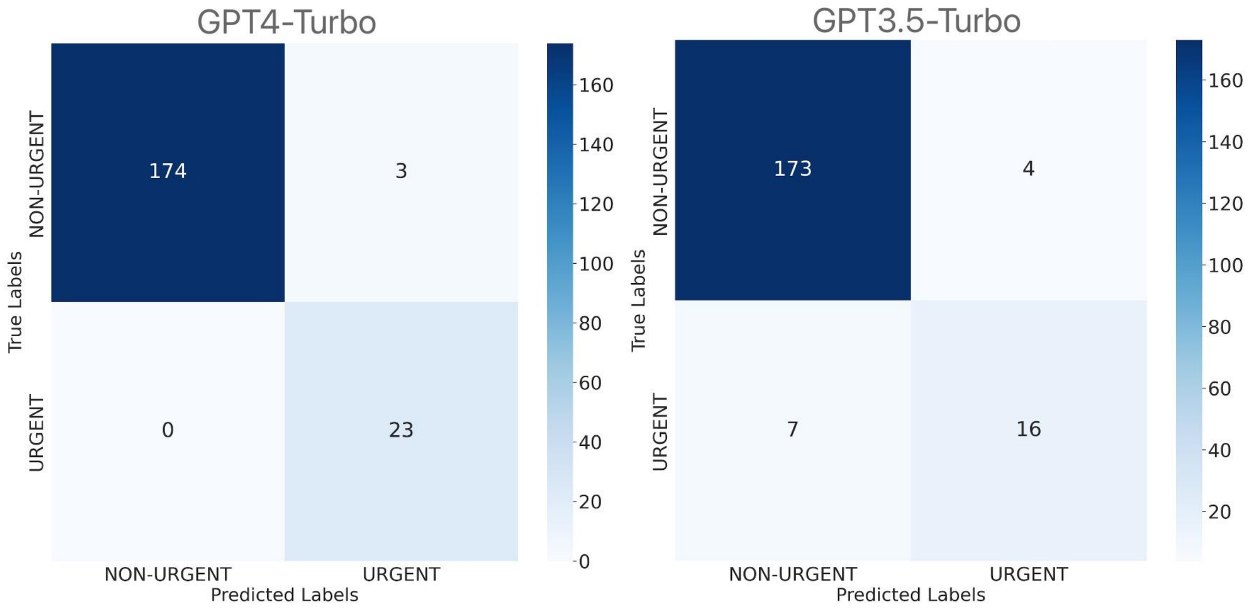


Figure 4. Validation Confusion Matrices of GPT4-Turbo models (left) vs. GPT3.5-Turbo baseline models (right) on Summarization

Model	**Precision**	**Recall**	**F1 Score**
GPT4-Turbo	0.885	1.00	0.939
GPT3.5-Turbo	0.800	0.696	0.744

Table 7: Validation Performance Metrics of GPT4-Turbo models (left) vs. GPT3.5-Turbo baseline models (right) on Summarization

		Misclassified Real Email	**Misclassified Synthetic Email**
Summarizer	**GPT3.5-Turbo Validation**	1	10
	GPT4-Turbo Validation	2	1

Table 8: Real vs. Synthetic Data Performance on Summrizer

7.3 Hold-out Results

After confirming the validation results, we tested both models on 50 hold-out samples. The Classifier exhibited exceptional performance on this dataset, as depicted in Figure 5, achieving a perfect Macro-F1 score. Meanwhile, the Summarizer only encountered one false positive from a real email, resulting in an F1 score of 0.91. All summaries effectively concluded the intention with correct key details. These outcomes are consistent with our validation results, reinforcing the models' ability to generalize effectively to unseen data.

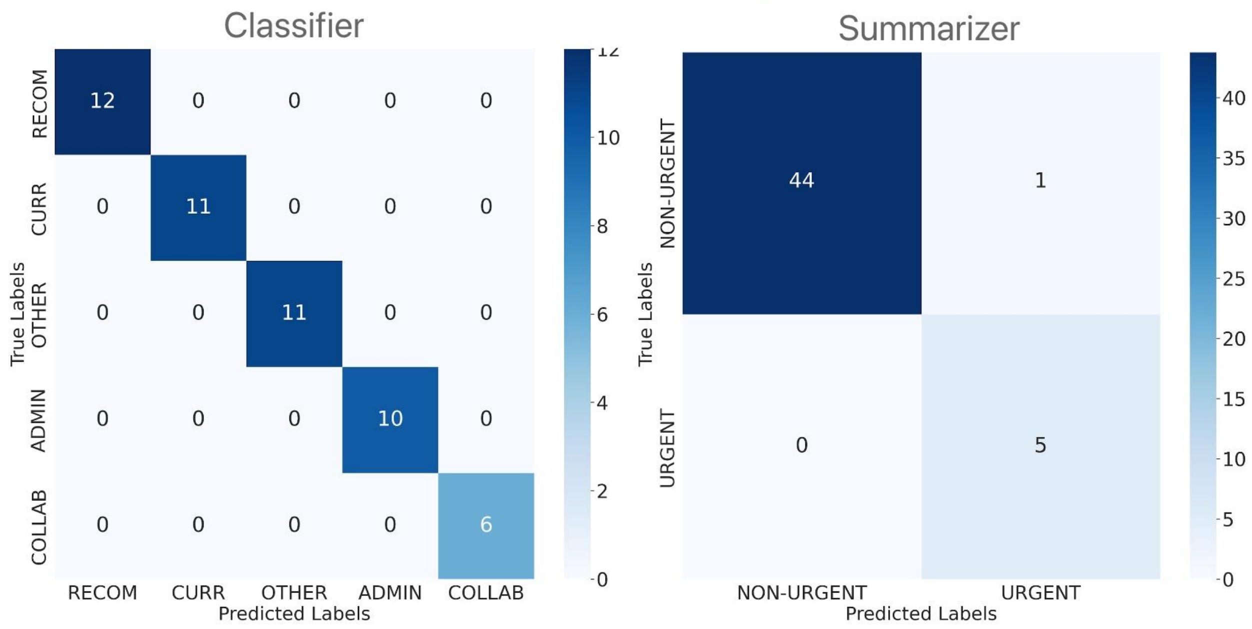


Figure 5. Confusion Matrices on Hold-out Data's Classification (left) and Summarization (right)

8.0 Qualitative Results

In this section, we delve into the qualitative performance of the models through examples of inputs, outputs, and chain-of-thought explanations.

8.1 Classifier Qualitative Results

After reviewing the explanation associated with each classification, we found that the explanations were precise for correct classifications, but in cases of misclassification, its logical analysis fell short of accurately grasping the conveyed information. For example, in one of the only two misclassifications, the model erroneously categorized a non-synthetic email from the COLLAB category as OTHER. This email, a thank-you note from a clinic director to a professor, was regarding a joint project on application development. Despite its relevance to university collaborations, the model misjudged it as unrelated to academic activities, likely due to its informal tone and content not explicitly mentioning the university.

The baseline model also misclassified the same email but labelled it as ADMIN instead of OTHER. This misclassification occurred despite the model recognizing the email's true intent. This suggests a misunderstanding of the ADMIN category by the baseline, aligning with the low precision observed for this category in Section 7.1.

8.2 Summarizer Qualitative Results

In the process of manually reviewing the summarizer results, we found that other than efficiently capturing and summarizing the urgency, the summarizer is also capable of capturing the mood which the author expressed in the email, through the tone of voice, the context and the wording of the emails. In contrast, the baseline model omits the urgency information if they are not explicitly expressed.

This is especially pronounced in a real email where Isabella was asking for an extension one day before the deadline. She was anxious and expecting a reply as soon as possible, but the expression did not include any urgent words. The summarizer correctly labelled this email as urgent and generated a concise summary of the email's emphasis. In the chain of thought explanation, the mood of the author was predicted accurately. However, the baseline misclassified the same email as non-urgent, as expected.

9.0 Discussion & Learnings

EmailSense, featuring four GPT4-Turbo engines - the dataset generator, urgency injector, classifier, and summarizer - has demonstrated outstanding performance. The dataset generator proficiently creates diverse, realistic email datasets, and the urgency injector skillfully integrates time-sensitive elements into emails. The high quality of the synthetic emails is particularly demonstrated in the baseline analysis where the majority of misclassifications stemmed from synthetic emails. This behaviour is consistent with their prevalence in the dataset (as there are more synthetic than real emails), underscoring their resemblance to real-world examples.

The integration of GPT4-Turbo into our email processing system showcases its remarkable capabilities in contextual understanding. Notably, the classifier achieves a 99% accuracy rate and a Macro F1 score of 0.990, and the summarizer excels in identifying urgent messages and producing concise, pertinent summaries, significantly surpassing the baseline model. However, the project's urgency detection feature could have been further enhanced by including the time and date of email receipt in the inputs to the classifier and summarizer. This addition is crucial as it compensates for GPT-4's limitation in accessing real-time data, improving its ability to assess urgency more accurately.

10.0 Individual Contribution

Isabella Hao:

- Collected emails in OTHER, ADMIN, and RECOM
- Wrote and evolved the prompt for the classifier and summarizer
- Implemented Summarizer with Weizhou
- Ran classifier and summarizer on the validation dataset
- Manually reviewed the classifier and summarizer results

Weizhou Wang:

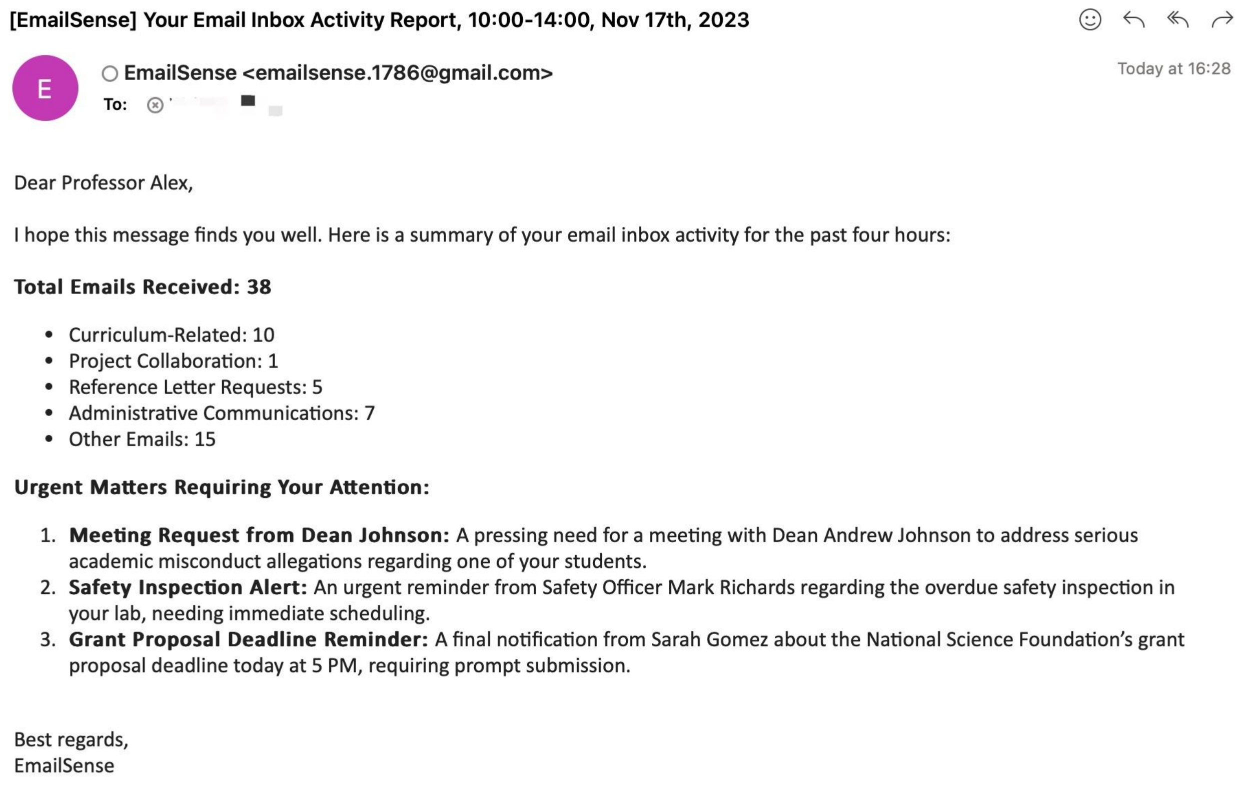
- Collected & Generated emails in COLLAB, and CURR categories
- Injected urgency to all emails
- Implemented Apple Shortcuts, Email Monitor Engine, and Result Generator
- Implemented Classifier with Isabella
- Tested baseline models’ performance

11.0 References:

[1] J. DeMers, “Email productivity benchmark report (May 2023),” Email Analytics: Visualize your team’s email activity (Gmail & Outlook), https://emailanalytics.com/email-productivity-benchmark-report/ (accessed Oct. 29, 2023).
[2] T. Costa, “How to email your professor in college,” Southern Utah University, https://www.suu.edu/blog/2019/10/email-your-professor.html (accessed Oct. 29, 2023).
[3] C. Nattrass, Etiquette tips for emailing your professors, https://nattrass.utk.edu/Etiquette.html#:~:text=However%2C%20their%20estimates%20of%20how,emails%2Fday%20about%20class%20alone.&amp;text=This%20is%20generally%20a%20good,receive%20%22thank%20you%22%20emails (accessed Oct. 29, 2023).
[4] M. Heiding, B. Schneier, A. Vishwanath, and J. Bernstein, Devising and Detecting Phishing: large language models vs. Smaller Human Models, Aug. 2023. Accessed: 2023. [Online]. Available: https://www.researchgate.net/publication/373332863_Devising_and_Detecting_Phishing_large_language_models_vs_Smaller_Human_Models
[5] S. Zhang, A. Celikyilmaz, J. Gao, and M. Bansal, “Emailsum: Abstractive email thread summarization,” arXiv.org, https://arxiv.org/abs/2107.14691 (accessed Oct. 29, 2023).

12.0 Appendices

Appendix A. Example Output from EmailSense



[EmailSense] Your Email Inbox Activity Report, 10:00-14:00, Nov 17th, 2023

○ EmailSense <emailsense.1786@gmail.com>

To: [redacted]

Today at 16:28

Dear Professor Alex,

I hope this message finds you well. Here is a summary of your email inbox activity for the past four hours:

Total Emails Received: 38

- Curriculum-Related: 10
- Project Collaboration: 1
- Reference Letter Requests: 5
- Administrative Communications: 7
- Other Emails: 15

Urgent Matters Requiring Your Attention:

1. **Meeting Request from Dean Johnson:** A pressing need for a meeting with Dean Andrew Johnson to address serious academic misconduct allegations regarding one of your students.
2. **Safety Inspection Alert:** An urgent reminder from Safety Officer Mark Richards regarding the overdue safety inspection in your lab, needing immediate scheduling.
3. **Grant Proposal Deadline Reminder:** A final notification from Sarah Gomez about the National Science Foundation’s grant proposal deadline today at 5 PM, requiring prompt submission.

Best regards,
EmailSense