Efficient Localized Inbox Filtering using Advanced Language Models: Preserving Privacy and Speed

Raz Gaon
Yuval Mamana

6.5940 - TinyML and Efficient Deep Learning Computing
Massachusetts Institute of Technology

December 13, 2023

Abstract

This paper introduces a novel email management system that harnesses the power of a fine-tuned Large Language Model (Mistral) (Jiang et al.) for efficient email summarization and organization. Central to our approach is a custom Gmail client that captures emails and processes them locally, ensuring user privacy and data security. The system employs the LLM to generate concise summaries of each email, stored in a local database, significantly reducing information overload. A personalized filtering algorithm, adaptable to user preferences, categorizes emails to align with individual needs. Additionally, the system compiles markdown reports of prioritized emails, offering a clear overview of important communications. This research highlights the application of machine learning in improving digital communication efficiency while addressing privacy concerns, marking a significant step towards intelligent and user-centric email management solutions.

1 Introduction

In the rapidly evolving domain of digital communication, managing the ever-increasing volume of emails remains a significant challenge for individuals and organizations alike. This paper introduces an innovative email management system that leverages the advancements in large language models to address this challenge. Our system not only enhances inbox organization and content prioritization but also underscores the importance of user privacy and data security in its design.

Central to our approach is the integration of a custom Gmail client with a state-of-the-art, fine-tuned Mistral LLM - Intel Neural Chat. This integration enables the automatic capture and summarization of incoming emails, effectively condensing the noisy information into manageable, concise summaries. These summaries are stored locally in a user-specific database, emphasizing our commitment to minimizing reliance on external cloud servers, thereby enhancing data privacy and processing speed.

A distinguishing feature of our system is its adaptive filtering mechanism. By harnessing the flexibility of System Messages, the system dynamically aligns with individual user preferences, categorizing emails based on their relevance and importance. This personalized filtering is not static; it evolves with the user's changing needs and preferences, showcasing the adaptability and learning capabilities of our model.

Moreover, our system's capability to generate structured markdown reports of filtered emails offers users a clear and organized overview of their essential communications. This feature is particularly beneficial for managing high-volume email traffic, providing a streamlined and efficient way to access prioritized information.

This work contributes to the field by demonstrating the practical application of large language models in managing daily digital tasks, especially in an era where data privacy concerns are paramount. Our system not only aids in efficient email management but also serves as a blueprint for future developments in privacy-conscious, adaptable, and user-centric digital communication tools. Through this research, we aim to set a new standard in email management, paving the way for more intelligent, efficient, and privacy-aware solutions in the realm of digital communication.

2 Methods

This section delineates the architecture and design principles of our novel email management system, which incorporates a fine-tuned Mistral LLM for email summarization and organization. The system is implemented as a Python library, adhering to abstract design patterns to ensure modularity and flexibility in swapping components like the LLM provider or email service provider.

2.1 System Architecture

Our system is built around a central Python library that orchestrates the various components. The architecture is divided into four main modules:

- **Email Capture Module:** Utilizes a custom Gmail client to fetch emails. This module is designed with abstract interfaces, allowing for easy integration with other email service providers if needed.
- **Summarization Module:** Integrates with the Neural Chat LLM to summarize the content of each email. The LLM component is also designed

to be replaceable, enabling the use of different language models in the future without significant codebase changes. Unfortunately, we faced performance issues when running Streaming LLM, but it can easily be used by changing two lines of code.

- **User Preference Filtering Module:** Filters emails based on user-defined preferences. This module adapts to the evolving preferences of the user, ensuring relevance and personalization in email categorization.
- **Report Generation Module:** Compiles summaries of emails that pass the user preference filter into a markdown report, providing a clear and organized overview of prioritized emails.

2.2 Abstract Design Patterns

To facilitate flexibility and scalability, our system employs abstract design patterns. This approach allows for the decoupling of specific functionalities from their implementations, making it straightforward to substitute components like the LLM or email providers.

2.3 Custom Database Class

For storing email summaries and user preferences, we developed a custom database class that interacts with JSON files. This choice was driven by the need for simplicity and ease of data manipulation. The JSON-based database allows for quick retrieval and update of data, catering to the dynamic nature of email communication and user preferences. The database class is equipped with methods for efficient querying, updating, and storing of data, all while ensuring data integrity and consistency.

2.4 Implementation Details

The entire system is implemented in Python, chosen for its rich ecosystem of libraries and its suitability for rapid prototyping and machine learning applications. The use of Python also facilitates easy integration with various machine learning models and email APIs. The system's modular design ensures that it remains adaptable and maintainable, allowing for future enhancements and integration with emerging technologies.

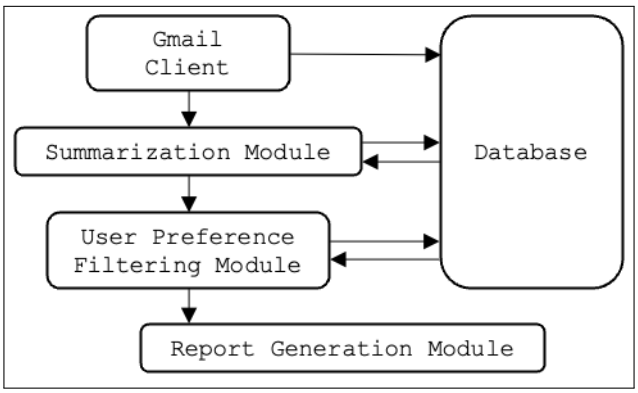

Figure 1: System architecture flowchart

3 Results

Our experimental results demonstrate the efficacy and efficiency of the proposed email management system. The system was evaluated in terms of processing time, accuracy in summarization, and effectiveness in filtering emails based on user preferences.

3.1 Processing Time

The system was tasked with processing a batch of 100 emails. The end-to-end processing, which includes capturing, summarizing, filtering, and report generation, was completed in approximately 3 minutes. This performance metric highlights the system's capability to handle substantial email volumes efficiently, making it suitable for users dealing with high-traffic inboxes.

3.2 Summarization and Filtering Accuracy with Llama 2

Initial experiments were conducted using Llama 2 (Touvron et al.) as the underlying language model for email summarization. It was observed that 50% of the emails that passed through the filtering process did not actually meet the defined user preferences. Further investigation revealed that the core issue lay in the quality of the summaries produced by Llama 2. These summaries often failed to capture the essential content of the emails accurately, leading to ineffective filtering.

3.3 Noise in Emails

An additional challenge identified was the presence of excessive noise in the emails, such as emojis and superfluous whitespace, which further hindered accurate summarization and subsequent filtering.

3.4 Experiments with Neural-Chat

In response to these challenges, we conducted further experiments using Neural-Chat (published by Intel), a fine-tuned version of the Mistral LLM. The results showed a significant improvement in the system's performance. With Neural-Chat, only 15% of the filtered emails did not align with the user's preferences, indicating a more accurate capture of email essence and effective filtering. This improvement underscores the importance of selecting a suitable language model for the task of email summarization and highlights the potential of fine-tuned models like Neural-Chat in handling the nuances and complexities of natural language found in emails.

4 Conclusion

In this paper, we have presented a novel email management system that leverages advanced machine-learning techniques to streamline the processing and organization of emails. The system, developed as a Python library with abstract design patterns, offers flexibility, scalability, and a user-centric approach to email management. Our experimental results demonstrate the system's efficiency in processing large volumes of emails and the importance of the choice of language model in determining the accuracy of email summarization and filtering.

Through our experimentation, we observed that while the initial use of Llama 2 for email summarization resulted in a significant portion of emails failing to meet user preferences, the shift to Neural-Chat, a fine-tuned version of the Mistral LLM, led to a marked improvement. This transition highlighted the critical role of a well-suited language model in handling the complexities and nuances of email content.

Our system addresses not only the efficiency of email management but also the critical aspects of user privacy and data security by ensuring local processing of data. The modular design allows for easy adaptation and integration with various email and LLM providers, showcasing the system's flexibility and forward compatibility.

Future work could explore the integration of more sophisticated filtering algorithms, the ability to handle a wider range of email formats, and the application of the system in different contexts beyond email management. Additionally, further refinement of the language model could be investigated to enhance the accuracy of summarization and filtering.

In conclusion, our research contributes to the field of digital communication by providing an efficient, user-friendly, and privacy-conscious solution to email management, demonstrating the potential of machine learning techniques in transforming everyday tasks.

References

[1] Jiang, Albert Q., et al. Mistral 7B. arXiv:2310.06825, arXiv, 10 Oct. 2023. arXiv.org, https://doi.org/10.48550/arXiv.2310.06825.

[2] Touvron, Hugo, et al. Llama 2: Open Foundation and Fine-Tuned Chat Models. arXiv:2307.09288, arXiv, 19 July 2023. arXiv.org, https://doi.org/10.48550/arXiv.2307.09288.

[3] Intel/Neural-Chat-7b-v3-1 · Hugging Face. https://huggingface.co/Intel/neural-chat-7b-v3-1. Accessed 11 Dec. 2023.