Chat Summary Model

Problem Description

When using group chats important information can get lost in the flow of messages. This problem gets compounded when users aren’t able to be involved for every message. It can take anywhere from 5-15 minutes or more for a user to catch up on unread messages depending on how active the group message is, especially if the group message is filled with detailed technical jargon or many important messages. According to a 2023 survey in the US and UK, 40% of the responders reported being overwhelmed by the high volume of group chat messages and notifications. There are benefits of being able to quickly summarize group chat messages as well as being able to highlight important information.

In my personal experience, corporate group chats can sometimes get filled with many messages that are only relevant during the time the chat is actively happening. There were many times where the discussion around the actual important information makes it difficult to get the important info, where multiple people asking clarifying questions can almost bury the truly important messages. If there was a way to filter through the messages automatically and give a summary of the message it would have been very helpful.

Impact Assessment

Information overload directly undermines user experience by creating cognitive fatigue and frustration. In an extreme example, coming back to a work chat after missing a day with over 100 messages can take between 30 minutes to an hour to read through. Of course this depends on how dense the messages are, but glancing through the messages can lead to missing important information. This can lead to users thinking that the messaging service is messy and unimportant to read every message. If there isn't an efficient way to receive a summary of the key points in the chat, the users may end up not using the chat service.

This would lead to users muting notifications, leaving groups, changing messaging providers, or limiting participation in on-going discussions. Long term this can weaken the effectiveness of the group chat, where many users are ignoring the chat and not voicing their opinions on what is being discussed. The more users that become disengaged, the less varied opinions are brought forward. This leads to stagnation and can cause users who have ideas to not contribute them due to simply not paying attention to a chat that to them seems like it isn’t worth their time to interact with due to constantly feeling out of the loop.

Platforms that fail to address information overload risk losing users to outside applications and messaging services. Platforms like Slack and Microsoft Teams have already integrated AI driven summarization or “smart recap” features, and if the internally developed system doesn’t have these features the employees may end up taking their conversations off the company approved platform to have access to these features. In order to retain enterprise and professional clients to the service, Acme Communications needs to begin integrating these features.

Proposed Solution:

Adding an Automated Dialogue Summarization Feature offers a direct, scalable solution to these challenges. By leveraging natural language processing (NLP) and transformer based models, the AI based system can condense these overwhelming group conversations into concise, coherent summaries that capture the key points. Summarizing the key decisions, updates, and next steps for the users as well as addressing the multiple pain points for the users and the businesses.

Primarily it will reduce cognitive load on the users. The summarization will enhance accessibility for users who may be missing the discussions due to being in a different time zone, having a busy day, or not actively checking the chat. The AI powered summaries will be able to provide a streamlined way of catching up for these users, allowing the process to take significantly less time than before. This will minimize time and effort needed to stay engaged in the chat, creating a more user friendly communication experience that encourages engagement.

The summaries also will act as a tool to allow users to quickly rejoin a discussion without losing context. This is especially valuable for companies using the service with teams in multiple countries, large teams, and customer support groups, where maintaining flow of conversation is essential. It also can support multilingual accessibility by integrating translation and abstraction capabilities, ensuring all users can understand the core messages regardless of language differences.

Adding these features will also allow Acme to obtain a higher industry value. The company can rebrand from just being a communication tool to an intelligent collaboration platform. Adding the summarization feature also aligns with Acme’s goal of embedding AI into the user experience, reinforcing the brand’s reputation for innovation and productivity. This also opens the door to offering different tiers of end user packages, such as advanced summarization tiers, searchable summary archives, and improved integration with productivity tools. These tiers could offer multiple levels of user experience and open the doors to different levels of end users to subscribe to the platform, increasing overall revenue.

Success Metrics:

Summarization Quality:

Model effectiveness for summarization will be measured with the industry-standard ROUGE(Recall-Oriented Understudy for Gisting Evaluation) scores, which measure the overlap between machine-generated and human-written summaries. Target ROUGE-1 (unigram overlap): >= .45, target ROUGE-2 (bigram overlap): >= .25, and target ROUGE-L (longest common subsequence): >= .40. These thresholds are consistent with strong performance benchmarks reported in dialogue summarization research and will indicate that the model is capturing both key details and contextual flow effectively.

User Engagement:

To measure success in user engagement, the benchmarks are: Reduce average catch-up time by 50%, decrease missed information reports by at least 30%, increase daily active engagement by 10-15%, and increase Net Promoter Score (NPS) by 10 points. These metrics will show that the model not only improved efficiency but also improved user satisfaction and retention.

Technical Performance Requirements:

In terms of Technical performance requirements, some good benchmarks initially would be: an average response time of =< 10 seconds, memory usage below 1gb of usage, and being able to handle 100-500 users concurrently. It would be better to have lower response time and concurrent user ability, but to begin those benchmarks would be acceptable to prove the system works.

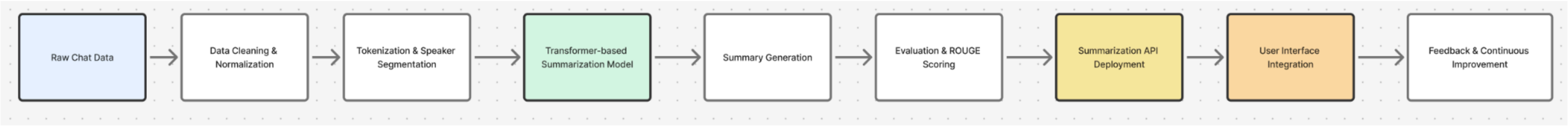
Development Process:

1. Data exploration and prep:
    a. Obtain and examine SAMSun dataset
    b. Clean and normalize the data
    c. Annotation alignment to ensure the summaries align with the reference
    d. Data augmentation using paraphrasing
2. Model architecture selection
    a. Select an appropriate transformer based architecture, like BART or T5
3. Model fine-tuning and optimization
    a. Hyperparameter tuning for optimal ROUGE performance
    b. Model compression to reduce model size
    c. Prompt-based fine-tuning to emphasize summarization goals
4. Evaluation and testing

    a. ROUGE-1, ROUGE-2, ROUGE-L for overlap based performance comparison
    b. Human evaluation to test accuracy, coherence, and usefulness
    c. A/B testing to test satisfaction and engagement with and without summarization included.

5. Prototype integration and user testing
    a. Once model is validated a working prototype should be integrated to ACME's chat interface to collect real world performance data with new data
6. Deployment and scaling considerations
    a. Focus on reliability, scalability, and cost efficiency
    b. Gradually implement features like containerization and batching to reduce load
7. Continuous improvement
    a. After receiving user input and feedback work on performing iterative improvements to improve the model
    b. Add features like topic-specific summarization and sentiment extraction to better enhance user experience

Flowchart for data processing:


Raw Chat Data
Data Cleaning & Normalization
Tokenization & Speaker Segmentation
Transformer-based Summarization Model
Summary Generation
Evaluation & ROUGE Scoring
Summarization API Deployment
User Interface Integration
Feedback & Continuous Improvement

Model and evaluation rationale:

A BERT-Based encoder, like BART or T5, is well suited for dialogue summarization because it combines robust contextual understanding and fluent text generation. The encoder captures bidirectional context from the dialogue, allowing the model to understand dependencies across speakers, tone shifts, and conversational flow. The decoder generates coherent, abstractive summaries rather than merely extracting sentences. Compared to unidirectional models, BERT-style encoders provide a richer representation of meaning, essential for handling informal, fragmented, or multi-speaker dialogues typical with messaging data. This architecture also enables the system to generate human-like summaries.

Fine tuning a pre-trained transformer offers substantial efficiency and performance benefits. It primarily reduces computational cost by leveraging existing linguistic knowledge which drastically reduces training time. A pretrained model also already understands grammar, semantics, and discourse structure, allowing the summarization task to adapt quickly with a limited dataset. The SAMSum dataset is relatively small, making it crucial to apply a pretrained model to achieve high ROUGE scores. Using a pretrained model also allows the model to be fine tuned to specific chat types such as corporate chat, customer support, or community management contexts by simply obtaining a dataset that matches with the use type needed.

ROUGE is the industry standard for dialogue summarization, making it the best metric to evaluate by. The different segments (ROUGE-1, ROUGE-2, and ROUGE-L) all capture different levels of abstraction in the summarization, and coupled with human evaluation for readability, coherence, factual accuracy, and usefulness the testing should both perform well and show good user experience. This combination of quantitative (ROUGE) and qualitative (Human) will provide a balanced assessment of real world summarization quality.

Optimizing dialogue summarization models requires balancing linguistic accuracy and computational efficiency. The learning rate will gradually be reduced to stabilize fine-tuning and prevent catastrophic forgetting. Then monitoring validation ROUGE to halt training before overfitting through early stopping will be used. Gradient accumulation will be applied to allow effective training on longer sequences without exceeding GPU memory limits. Knowledge distillation compresses large transformer models into smaller, faster ones while retaining semantic accuracy, crucial for deployment stability. Finally Through quantization and pruning the model redundant weights will be removed to enhance inference speed without major quality loss. Together these techniques ensures the summarization model remains accurate, responsive, and scalable, a critical balance for real-time messaging applications.

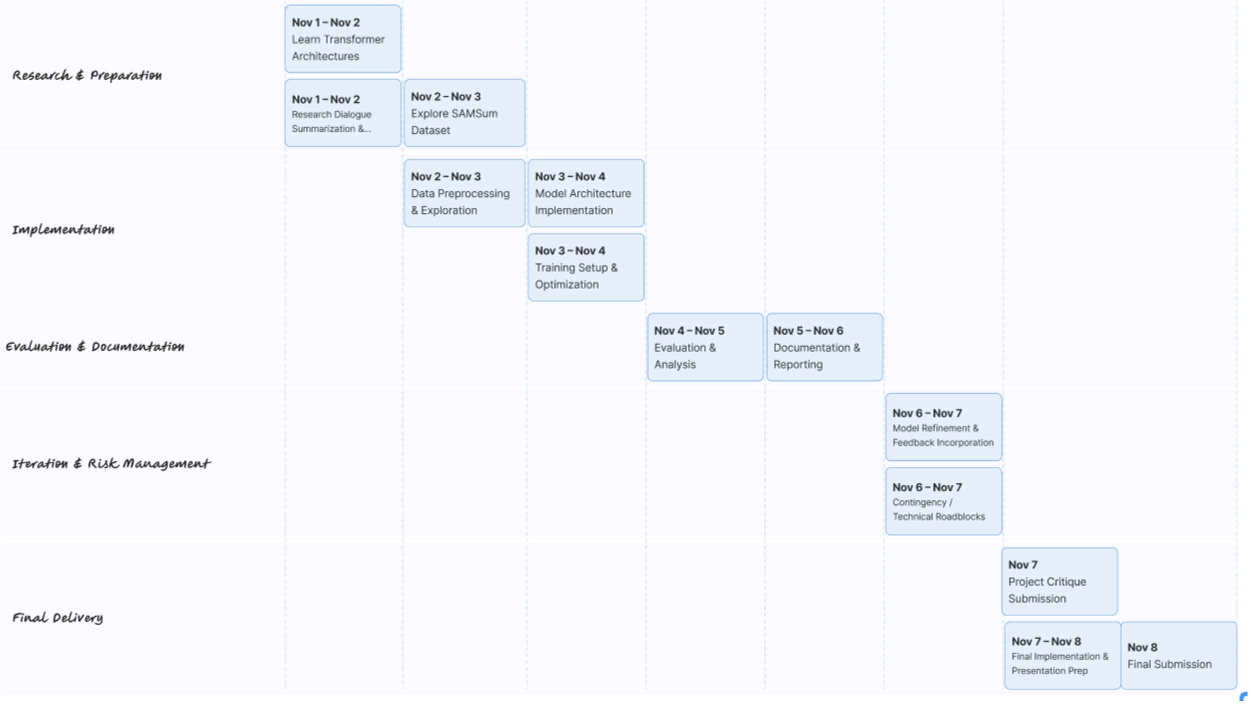
Alignment with requirements:

The proposed dialogue summarization solution effectively meets both business and technical goals. It fulfills all project deliverables—from problem definition and data processing to model design, evaluation, and deployment—while remaining feasible and scalable. By fine-tuning a BERT-based encoder-decoder model, the system achieves high-quality summaries without incurring the computational cost of training from scratch.

From a business standpoint, it directly addresses Acme’s need to reduce information overload and enhance user engagement. The model generates concise, meaningful summaries that help users stay informed efficiently, improving satisfaction and retention. Moreover, its lightweight deployment and optimization techniques ensure fast, cost-effective operation. Overall, this approach balances advanced NLP performance with real-world usability, demonstrating clear strategic value for Acme’s communication platform.

Timeline:

Developed using Figma:



Research & Preparation	Nov 1 – Nov 2 Learn Transformer Architectures							
	Nov 1 – Nov 2 Research Dialogue Summarization &...	Nov 2 – Nov 3 Explore SAMSum Dataset						
Implementation		Nov 2 – Nov 3 Data Preprocessing & Exploration	Nov 3 – Nov 4 Model Architecture Implementation					
			Nov 3 – Nov 4 Training Setup & Optimization					
Evaluation & Documentation				Nov 4 – Nov 5 Evaluation & Analysis	Nov 5 – Nov 6 Documentation & Reporting			
Iteration & Risk Management						Nov 6 – Nov 7 Model Refinement & Feedback Incorporation		
						Nov 6 – Nov 7 Contingency / Technical Roadblocks		
Final Delivery							Nov 7 Project Critique Submission	
							Nov 7 – Nov 8 Final Implementation & Presentation Prep	Nov 8 Final Submission